

VĚDECKÉ SPISY VYSOKÉHO UČENÍ TECHNICKÉHO V BRNĚ

Edice Habilitační a inaugurační spisy, sv. 808

ISSN 1213-418X

Martin Čadík

**VISUAL GEO-LOCALIZATION
AND CAMERA POSE ESTIMATION
IN NATURAL ENVIRONMENTS**

BRNO UNIVERSITY OF TECHNOLOGY
Faculty of Information Technology
Department of Computer Graphics and Multimedia

doc. Ing. Martin Čadík, Ph.D.

**VISUAL GEO-LOCALIZATION
AND CAMERA POSE ESTIMATION
IN NATURAL ENVIRONMENTS**

**VIZUÁLNÍ GEOLOKALIZACE
A ODHAD PÓZY KAMERY V PŘÍRODNÍM PROSTŘEDÍ**

**TEZE PŘEDNÁŠKY
K PROFESORSKÉMU JMENOVACÍMU ŘÍZENÍ
V OBORU VÝPOČETNÍ TECHNIKA A INFORMATIKA**



BRNO 2025

KEYWORDS

visual localization, vision based localization, geo-localization, image to model registration, terrainaided navigation, cross-domain registration, camera pose estimation, camera orientation estimation, extrinsic calibration, photograph peak tagging, automatic geo-registration, place recognition

KLÍČOVÁ SLOVA

vizuální lokalizace, lokalizace založená na vizuální informaci, geo-lokalizace, registrace obrazu a modelu, navigace s pomocí terénu, registrace napříč doménami, odhad pozice kamery, odhad orientace kamery, kalibrace kamery, označení vrcholů na fotografii, automatická georegistrace, vizuální rozpoznání místa

Contents

About the author	4
1 Introduction	5
1.1 Motivation, scope and structure	5
1.2 Acknowledgments	7
2 Problem statement and related work	8
3 Camera orientation estimation	11
3.1 Camera orientation estimation using 3D terrain models	11
3.2 Camera pose estimation using learned descriptors	13
3.3 Camera pose estimation from line correspondences	14
3.4 Camera orientation estimation using cascaded attention	15
4 Visual geo-localization	17
4.1 New datasets for visual localization	17
4.1.1 Dataset GeoPose3K	17
4.1.2 Dataset Alps100K	18
4.1.3 Dataset CrossLocate	19
4.2 Skyline detection for visual localization	20
4.3 Novel geo-localization methods	20
4.3.1 Geo-localization using geo-registration	21
4.3.2 Cross modal retrieval-like geo-localization	22
5 Applications	24
5.1 Automatic label placement	24
5.2 Monocular depth estimation	26
5.2.1 3D model-based synthesis of outdoor depth maps	26
5.2.2 Neural synthesis of outdoor depth maps	27
6 Conclusions and future work	30
References	31
Abstract	38
Abstrakt	38

About the author

Martin Čadík earned his Master's degree in Computer Science and Engineering from the Faculty of Electrical Engineering, Czech Technical University (CTU) in Prague, in 2002, graduating with a "red diploma." In 2008, he completed his Ph.D. studies at the same faculty with honors. His dissertation [Čadík, 2008b] was awarded an Outstanding Dissertation Award.

From 2007 to 2009, he worked as a researcher in the Computer Graphics Group at the Faculty of Electrical Engineering, CTU in Prague. Between 2009 and 2013, he served as a postdoctoral researcher at the Max-Planck-Institut für Informatik in Saarbrücken, Germany, collaborating with Karol Myszkowski, Hans-Peter Seidel, Lionel Baboud, and Elmar Eisemann on perceptual issues in computer graphics, computer vision, and image processing. In 2013, he joined the Department of Computer Graphics and Multimedia at Brno University of Technology (BUT) in Brno, Czech Republic, as an assistant professor. In the same year, he founded the CPhoto@FIT research group at BUT, where he has served as its head and executive director. In 2015, he defended his habilitation thesis [Čadík, 2015] and was promoted to associate professor at BUT.



Martin Čadík has conducted research visits and stays at several prestigious institutions. In 2004, he collaborated with Alessandro Artusi and Michael Wimmer at the Institute of Computer Graphics and Algorithms, Vienna University of Technology, Austria, on tone mapping algorithms. In 2005, he worked with Attila Neumann, Stefan Jeschke, and Michael Wimmer on high dynamic range (HDR) imaging at the same institute. From 2005 to 2006, he joined the Department of Informatics at Hochschule für Technik und Wirtschaft in Dresden, Germany, where he collaborated with Walter Paetzold and László Neumann on developing novel tone mapping methods. In 2007, he worked with Rafael Garcia and László Neumann at the Computer Vision and Robotics Group, University of Girona, Spain, on HDR tone mapping and color-to-grayscale conversions. In 2012, he visited Bangor University, Wales, UK, to work with Rafał Mantiuk on objective image quality assessment. In 2016, he collaborated with Josef Sivic at INRIA – Willow Project in Paris, France, on visual geo-localization. Most recently, in 2019, he visited Purdue University, USA, where he worked with Bedřich Beneš on computational photography and visual perception.

He has participated in several research grants, including: 2019: DEEP_LOCATE - Deep-Learning Approach to Topographical Image Analysis, MŠMT CZ, 2019-2022; 2018: BOREC - Colour Image in Real-time Embedded Computing (TH03010330, Technology Agency of the Czech Republic); 2016: Visual Geo-localization and Pose Estimation in Mountainous Terrains (project OPEN-7-49, National Supercomputing Center IT4Innovations, Czech Republic); 2014: Incoming Grant on Visual Localization in Natural Environments (SoMoPro II co-financed by Marie Curie Actions, REA 291782); 2012: COST STSM (ECOST-STSM-IC1005-100312-015783), Image distortion maps for evaluation of quality metrics, Bangor, UK; 2006–2010: Grant No. LC06008, Center for Computer Graphics; 2007–2008: Aktion Kontakt OE/CZ grant No. 48p11, Realistic Real-Time Rendering of Trees and Vegetation; 2005–2006: Erasmus scholarship of the European Union, HTW Dresden, Germany; 2007: Grant No. CTU0715413, Organization of the EG 2007; 2004: Grant No. CTU0408813, Automatic Comparison of Images.

Martin Čadík's research interests include visual geo-localization, high dynamic range imaging, image processing, computer vision, human visual perception, and image and video quality assessment. He is the author or co-author of more than 50 scientific publications in journals and conferences and holds two patents. His current SCOPUS h-index is 18. He regularly reviews for major conferences and journals in computer vision, graphics, and image processing. He has served as an associate editor for the *Computer Graphics Forum* journal and as a program committee member for several international conferences. In 2014, together with Jaroslav Krivánek, he co-founded the *HighVisualComputing* (hiviscomp.cz), which he continues to organize.

As a faculty member of BUT's Faculty of Information Technology, Martin Čadík is actively engaged in teaching. He supervises bachelor, master's, and Ph.D. students and teaches courses such as Computer Vision, Advanced Computer Graphics, Computer Art, Computational Photography, and Introduction to Game Development. He has introduced and co-developed novel courses on Computational Photography and Game Development. Additionally, he is active in popularizing computational photography through public media.

1 Introduction

Visual communication has become an essential aspect of modern life. The widespread use of digital cameras, smartphones and handheld devices equipped with built-in cameras has turned nearly everyone into a photographer. Each day, we capture an ever-increasing number of photographs and videos, often with enhanced technical quality. These visual assets are shared, edited, searched, archived, and enhanced; they serve specific purposes or are simply created to produce visually appealing content.

For personal photographs and videos, the approximate *position* of their capture is often known. However, for the majority of imagery available online today, this information is typically unknown. Even when the capture position is available, such as through GPS tags, the *orientation* of the camera is usually missing. The knowledge of both camera position and orientation has a vast range of important applications, including the organization of photo collections, environmental monitoring, heritage preservation, augmented reality, autonomous navigation, and uses in military, security, and rescue services.

In this work, we summarize our research on visual geo-localization and camera pose estimation. The primary objective of these methods is to determine the position and/or orientation of the camera using only the visual information encoded in the image. While this is an extremely challenging task, our results demonstrate that it is possible to successfully localize a significant number of images, and ongoing research is expected to further improve these results.

Our focus is particularly on *natural environments*, which present additional challenges for visual localization methods. Natural environments cover vast areas, are not densely represented by community photographs (unlike urban areas), and often exhibit self-similarity. Unlike urban settings, where structured and repetitive patterns such as buildings, roads, and street signs provide an abundance of localization cues, natural environments are characterized by irregular landscapes, vegetation, and dynamic lighting conditions. These scenes often lack distinguishable landmarks or man-made features, making reliable geo-localization and pose estimation particularly difficult.

1.1 Motivation, scope and structure

Knowing the location where an image or video was captured is valuable in numerous contexts. This is particularly true for *natural scenes*, where the location may be easily forgotten, even by the photographer. Additionally, understanding what is depicted in the image can be challenging, as natural scenes often contain few recognizable landmarks for non-experts. With the increasing number and quality of photographs being produced and shared today, the sheer volume of data becomes unmanageable without semantic information and automated techniques for organization. Determining the location of an image facilitates *semantic labeling*, which aids in *organizing and sharing photos*, as well as simplifying the *retrieval, archiving, and inspection* of large image collections [Luo et al., 2011].

Furthermore, having knowledge of the orientation of the camera opens the door to a wide range of advanced applications. Developments in *computational photography* leverage this information for innovative image enhancements, such as *dehazing, photo browsing, image relighting, and image-based modeling* [Kopf et al., 2008, Bae et al., 2010]. Applications in *artificial intelligence, augmented reality (AR), and mixed reality (MR)* also rely on accurate knowledge of camera pose. For example, these technologies can augment the visual field with additional information, metadata, or virtual objects to assist in cognitively demanding tasks, such as navigation in nature. An example from our work [Baboud et al., 2011] is illustrated in Fig. 1.

In *autonomous robot navigation*, these techniques enable robots and unmanned aerial vehicles (UAVs) to determine their location and orientation in complex environments, ensuring accurate path planning and obstacle avoidance without relying solely on GPS. In *surveillance and recognition*,

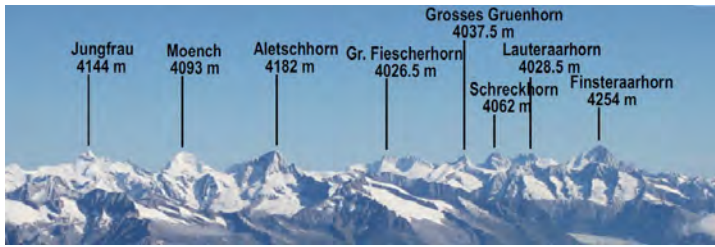


Figure 1: An application of geo-localization: when the camera pose is known, summit names can be overlaid onto the image, and other details such as peak heights can be inferred from geo-referenced 3D terrain models. This enhances the interpretability and usefulness of the captured panorama, providing valuable insights to the viewer.

visual localization assists in identifying key locations and monitoring specific regions, aiding security operations and situational awareness. *Environmental monitoring* also benefits significantly, as visual localization helps in tracking wildlife, assessing deforestation, and mapping natural disasters. Furthermore, in *heritage preservation*, these techniques facilitate the documentation and restoration of cultural landmarks by enabling precise localization of historical sites and integration with 3D modeling tools, ensuring that these treasures are preserved for future generations.

In *military, defense, and rescue operations*, visual localization and camera orientation estimation can enhance GPS readings, particularly when the GPS signal is compromised by occlusions, multipath errors, satellite and environmental disturbances, or device-specific magnetic interference. Additionally, GPS systems are vulnerable to deliberate attacks, such as GPS jamming (blocking the signal) or spoofing (sending false signals to deceive the receiver) [Haider and Khalid, 2016]. These vulnerabilities arise from the inherently weak GPS signals, which can be easily overridden. Such disruptions can severely impair GPS functionality, posing significant risks to navigation and positioning systems across various domains.

Visual geo-localization and camera pose estimation have garnered significant academic and public interest; however, these tasks remain exceptionally challenging and, in many respects, unresolved. The primary obstacles include limited and sparsely populated datasets, which are often biased toward notable landmarks and urban settings. Moreover, existing methods face accuracy constraints and are prone to errors caused by environmental variations, such as changes in lighting, weather, and scene dynamics. Our work seeks to address these challenges, reduce existing limitations, and stimulate further research in this critical domain.

This thesis focuses on *visual geo-localization and camera pose estimation* in natural environments, which is currently one of our primary research interests. In addition to this topic, our research spans several other areas that are not covered in this work, including:

- *Image and video quality assessment* [Čadík and Aydın, 2017, Čadík et al., 2013, Čadík et al., 2012, Herzog et al., 2012, Čadík et al., 2011, Aydın et al., 2010, Čadík and Slavík, 2004b, Čadík and Slavík, 2004a],
- *High dynamic range (HDR) image and video processing* [Příbyl et al., 2016, Čadík, 2015, Pajak et al., 2010b, Pajak et al., 2010a, Čadík et al., 2008, Čadík, 2007, Čadík et al., 2006, Fialka and Čadík, 2006, Čadík and Slavík, 2005],
- *Color-to-grayscale conversion* [Čadík, 2008a, Čadík, 2008b, Neumann et al., 2007],

- *Human perception in visual computing* [Rajasekaran et al., 2022, Polasek et al., 2021, Fišer et al., 2014, Sýkora et al., 2014, Sýkora et al., 2011, Aydin et al., 2010], and
- *Energy production forecasting* [Polasek and Čadík, 2023, Čadík et al., 2003].

This thesis is organized as follows. Chapter 2 provides a problem statement and a summary of prior work on visual geo-localization. Chapter 3 presents our contributions to camera orientation and pose estimation. In Chapter 4, we describe our advancements in visual geo-localization methods, while Chapter 5 explores new applications of these techniques. Finally, the thesis concludes with a discussion of future research directions.

1.2 Acknowledgments

First and foremost, I would like to express my heartfelt gratitude to all the members of my CPhoto@FIT research group and my esteemed colleagues who contributed to the papers presented in this thesis. In alphabetical order, my sincere thanks go to: Touqeer Ahmad, Lionel Baboud, Bedrich Benes, Petr Bobák, Jan Brejcha, Pavel Campr, Ladislav Čmolík, Shay Dekel, Stephen DiVerdi, Elmar Eisemann, Yannick Hold-Geoffroy, Zhili Chen, Ondřej Chum, Yosi Keller, Sungkil Lee, Michal Lukáč, Tomáš Polášek, Bronislav Příbyl, Jan Tomešek, and Oliver Wang.

I am also deeply grateful to Honza Černocký, Pavel Zemčík, and Adam Herout for their leadership of the Department of Computer Graphics and Multimedia at Brno University of Technology. Their support and encouragement have been invaluable in helping me complete this work. I would like to express my sincere gratitude to Karol Myszkowski and Hans-Peter Seidel, senior researcher and head of the Computer Graphics Group at the Max-Planck Institute for Informatics in Saarbrücken, respectively, for fostering an exceptionally stimulating environment for research and education, and for generously sharing their invaluable experience. I am also deeply thankful to Jiří Žára and Pavel Slavík for their leadership of the Department of Computer Graphics and Interaction at CTU in Prague and for their efforts in maintaining the smooth operation of the CG lab. I owe a debt of gratitude to the members of the CG lab at TU Vienna, including Werner Purgathofer, Michael Wimmer, Alessandro Artusi, and Attila Neumann, for their hospitality and support during my research visits to Vienna. My heartfelt thanks also go to Rafael García and László Neumann for their kindness and courtesy during my visits to UdG in Girona. I would be remiss not to mention Lionel Baboud, Robert Herzog, and Rafał Mantiuk, who, beyond being co-authors of my papers, are exceptional friends and colleagues. A special thanks to Jaroslav Křivánek—I cherish every moment I was fortunate to spend with you. Many others have influenced my work. I extend my thanks to all my colleagues and friends I have met in research groups in Brno, Prague, Pilsen, Vienna, Saarbrücken, Dresden, Zürich, Bangor, Girona, and Rennes for their valuable ideas, comments, and friendship. Finally, my deepest gratitude goes to my family, Hana and Anna, whose boundless patience and support made the completion of this work possible.

Parts of the work presented in this thesis were supported by the DEEP_LOCATE - Deep-Learning Approach to Topographical Image Analysis, MŠMT CZ, 2019-2022; the Incoming Grant on Visual Localization in Natural Environments (SoMoPro II co-financed by Marie Curie Actions, REA 291782); the MŠMT CZ, res. prog. No.: MSM-212300014, MSM-6840770014 (Research in the area of information technologies and communications); and LC-06008 (Center for Computer Graphics), by European Cooperation in Science and Technology EU RTD (ECOST-STSM-IC1005-100312-015783). Computational resources were primarily supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ ID:90140), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

2 Problem statement and related work

Visual geo-localization aims to estimate the geographical origin of a visual document, i.e., a photograph or a video, see Fig. 2. Visual localization has been researched for several decades [Brejcha and Čadík, 2017b, Hays and Efros, 2008, Weyand et al., 2016, Arandjelovic et al., 2016]. While tremendous progress could be observed in the localization focused on outdoor urban areas [Brachmann and Rother, 2018, Ge et al., 2020], localization targeted in nature remains an open problem. Existing methods localize with an error in order of kilometers at best [Saurer et al., 2016, Baatz et al., 2012], or assume a good initial pose estimate and only attempt to refine the camera pose [Baboud et al., 2011, Porzi et al., 2016a, Brejcha and Čadík, 2018].



Figure 2: The goal of visual geo-localization is to determine the position (x, y, z) of the camera (right) that captured the query image (left). Camera orientation estimation methods focus on calculating the camera’s rotation angles (α, β, γ) .

Natural environments introduce a number of specific challenges. They are highly variable, as both their appearance and geometry change with weather and seasons [Arroyo et al., 2016]. The amount of available data (i.e., user-taken photographs), is low compared to urban areas, and the spatial coverage is sparse and uneven. Accordingly, it is often impossible to localize in nature using only real photographs. Existing localization approaches, therefore, operate across multiple domains or image modalities, such as cross-view methods [Hu et al., 2018, Lin et al., 2013, Lin et al., 2015] utilizing satellite and aerial imagery or cross-modal methods utilizing 3D terrain models [Saurer et al., 2016].

Two significant aspects that influence the design of localization methods are target environment (*urban* [Arandjelovic et al., 2016, Wang et al., 2018, Kendall et al., 2015, Brahmbhatt et al., 2018, Torii et al., 2015, Armagan et al., 2017], *natural* [Saurer et al., 2016, Brejcha and Čadík, 2018, Baboud et al., 2011, Porzi et al., 2016a], *global* [Hays and Efros, 2008, Hays and Efros, 2015, Weyand et al., 2016, Vo et al., 2017]) and spatial scale (*city-scale* [Arandjelovic et al., 2016, Wang et al., 2018, Lin et al., 2015], *large-scale* [Saurer et al., 2016], *planet-scale* [Hays and Efros, 2008, Hays and Efros, 2015, Weyand et al., 2016, Vo et al., 2017]). The vast differences between the individual environments and scales lead to diverse approaches. As a result, many underlying localization principles may be observed (classification [Weyand et al., 2016, Gronát et al., 2016, Armagan et al., 2017], retrieval [Arandjelovic et al., 2016, Saurer et al., 2016, Hays and Efros, 2008, Hays and Efros, 2015, Vo et al., 2017, Torii et al., 2015, Radenović et al., 2016, Noh et al., 2017], regression [Kendall et al., 2015, Brachmann and Rother, 2018, Brahmbhatt et al., 2018], structure-from-motion [Hartley and Zisserman, 2004, Agarwal et al., 2009, Heinly et al., 2015, Geppert et al., 2019]).

Global planet-scale approaches [Hays and Efros, 2008, Hays and Efros, 2015, Weyand et al., 2016, Vo et al., 2017] attempt to localize images captured anywhere in the world, no matter the environment, which usually leads to localization errors in hundreds of kilometers. Therefore, these

methods may be useful for space pruning and scene type recognition. Specifically, PlaNet [Weyand et al., 2016] is a deep learning classification approach to geo-localization. The classification approach was later shown inferior to the image retrieval utilized by Revisited IM2GPS [Hays and Efros, 2008, Vo et al., 2017]. This highlights the advantage of building a general image descriptor in comparison to trying to memorize the entire world within a classification model. Furthermore, the retrieval approach required less training data while providing better performance. This is an important observation with respect to natural environments where data is extremely scarce.

Approaches aimed at outdoor (*sub*)urban environments [Torii et al., 2015, Armagan et al., 2017] are much more advanced and precise, as they have gained a lot of attention in recent years. They are typically used for *city-scale* localization [Arandjelovic et al., 2016, Wang et al., 2018, Lin et al., 2015], though some are precisely tuned for specific places or landmarks [Brachmann and Rother, 2018, Kendall et al., 2015]. While these *city-scale* approaches were designed for urban areas, they might represent a potential avenue to the solution of localization in nature. NetVLAD [Arandjelovic et al., 2016] successfully utilizes the retrieval approach. It combines custom feature aggregation with weakly supervised learning to perform place recognition despite changes in appearance over time. Unfortunately, the NetVLAD aggregation results in large descriptors, which are not ideal for our large-scale localization, even after the proposed dimensionality reduction to 4096 dimensions. However, a further reduction at the cost of accuracy might be possible. HOW [Tolias et al., 2020] is the state-of-the-art instance-level recognition (retrieval/localization) method trained on datasets of outdoor photographs of landmarks and buildings. It uses learned internal local descriptors (HOW) combined with an ASMK image search [Tolias et al., 2013] approach to perform search and classification in the domain of landmarks, where it outperforms existing global and local descriptors. To a certain extent, the ASMK is considered a replacement of traditional spatial verification. DELG [Cao et al., 2020] is another state-of-the-art large-scale image retrieval approach. Contrary to HOW, it utilizes both the global and local descriptors learned within a single model to perform two-step retrieval and instance-level recognition for outdoor landmark scenes.

Regression approaches to urban localization [Brachmann and Rother, 2018, Kendall et al., 2015] can provide sub-meter spatial precision, but they are usable only on small areas. Urban localization is often solved through structure-from-motion (SfM) [Hartley and Zisserman, 2004, Agarwal et al., 2009, Heinly et al., 2015]. SfM approaches [Irschara et al., 2009, Zeisl et al., 2015, Geppert et al., 2019] perform localization using 3D models acquired from many overlapping photographs. This requires millions of photographs, which makes SfM approaches unsuitable for large-scale localization in nature.

Methods focusing specifically on *natural* environments are far less explored. They often operate at a *large scale* [Saurer et al., 2016], corresponding to an area of a country or mountain range. A separate group of methods focuses only on camera pose refinement from a good initial estimate [Baboud et al., 2011, Porzi et al., 2016a]. *Global* and *urban* approaches typically utilize only ordinary photographs for localization. However, in *natural* environments, the number of user-taken photographs is low and the spatial coverage is sparse and uneven. This leads to various methods of cross-view or cross-modal character. The cross-view methods [Hu et al., 2018, Lin et al., 2013, Lin et al., 2015, Vo and Hays, 2016, Liu and Li, 2019, Shi et al., 2019] utilize databases of satellite or aerial imagery for the localization of ground-level queries. The cross-modal methods make use of digital elevation models to create databases of various synthetic modalities, such as skylines [Saurer et al., 2016]. Synthetic modalities (e.g., horizon lines and silhouette maps) were successfully used for a camera pose estimation in nature [Baboud et al., 2011, Porzi et al., 2016a]. Semantic segmentations were used for camera pose estimation both in nature [Brejcha and Čadík, 2018, Benbihi et al., 2020] and in a city [Armagan et al., 2017]. Assuming that localization methods are powerful enough to work across image domains, synthetic modalities might offer a solution to the localization in nature.

Horizon-based localization [Baatz et al., 2012, Saurer et al., 2016] (abbreviated HLoc) localizes photographs captured in mountains by extracting visible skylines and comparing them with synthetic horizon curves stored in a database. While skylines arguably carry useful information, other features might be more efficient, especially in situations where the horizon is obscured or not visible.

3 Camera orientation estimation

In this section, we present our work on camera orientation and pose estimation. Specifically, we introduce methods that leverage 3D terrain models to determine camera orientation in natural environments. One approach utilizes silhouette edge matching [Baboud et al., 2011], while another incorporates semantic cues to enhance the estimation process [Brejcha and Čadík, 2018]. We also demonstrate that camera pose can be estimated using line correspondences [Příbyl et al., 2015, Příbyl et al., 2017], providing an alternative approach for this task. Additionally, we demonstrate that feature-point-based matching is feasible in this context by employing learned descriptors [Brejcha et al., 2020]. Finally, we describe a novel approach for estimating extreme 3D image rotations using cascaded attention mechanisms [Dekel et al., 2024].

3.1 Camera orientation estimation using 3D terrain models

Given a photograph, our goal is to estimate its pose relatively to an accurate 3D terrain model based on a digital elevation map (DEM) [Baboud et al., 2011]. In this case, we assume that the camera’s field of view is known, as well as an estimate of the viewpoint position. Given these hypotheses, we are looking for the rotation that maps the camera frame to the frame of the terrain. The set of images that can be shot from the viewpoint is entirely defined by a spherical image centered at this viewpoint against which we need to match the query photo.

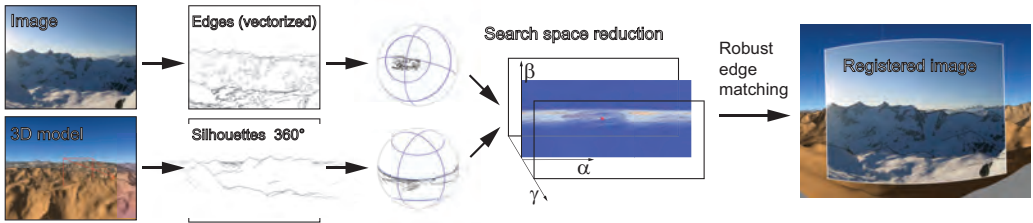


Figure 3: Overview of our camera orientation estimation method using silhouette edge matching. We leverage silhouette edges, which are among the most reliable features detectable in natural scenes. Using an edge detection algorithm, our technique identifies the best match between the detected silhouette edges and those rendered from a synthetic model. To address the inevitable noise in detected edges, we develop a robust matching metric that ensures reliable results. Additionally, we introduce a fast preprocessing step based on a novel spherical vector cross-correlation, which effectively reduces the search space and significantly accelerates the matching process.

We focus on outdoor scenes where relying on photogrammetry information is impractical due to its high variability. Instead, we rely on silhouette edges (see Fig. 3) that can be obtained easily from the terrain model and can be (partially) detected in the photograph. In general, the detected silhouette map can be error prone, but we enable a robust silhouette matching by introducing a novel metric. Our main observation relates to the topology of silhouette-maps: a feasible silhouette map in general configuration can contain T-junctions, but no crossings. Crossings appear only in singular views, when two distinct silhouette edges align. Consequently, a curve detected as an edge in the photograph, even if not silhouette, usually follows a feature of some object and thus never crosses a silhouette.

Because a direct extensive search using silhouette matching metric is costly, we additionally propose a fast preprocess based on novel spherical vector cross-correlation. It effectively reduces

the search space to a very narrow subset, to which the robust matching metric is then applied, see Fig. 3.

In the follow-up work [Brejcha and Čadík, 2018], we inquired whether areal information coming e.g. from semantic segmentation could improve the camera estimation performance. To this end, the terrain model is rendered with synthetic semantic segments as a spherical $360^\circ \times 180^\circ$ panorama (see Fig. 4(a)). A projective query image containing estimated semantic segments is projected on the unit sphere as well. The query image is scaled to cover the part of the unit sphere corresponding to its field-of-view.

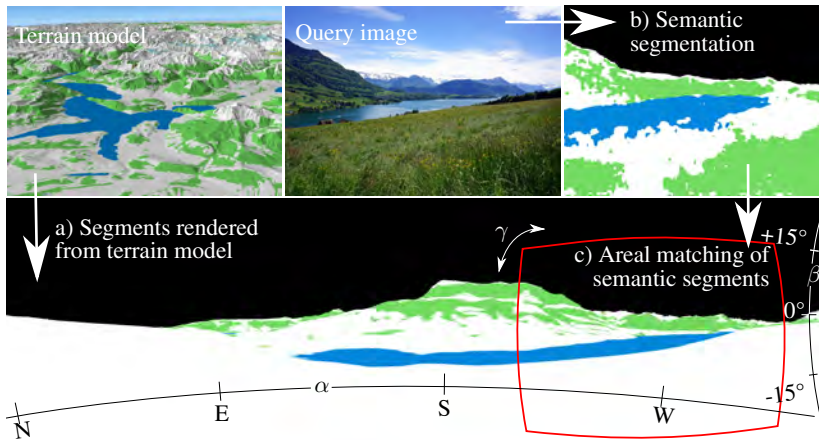


Figure 4: Overview of our camera orientation method using semantic cues. a) Synthetic semantic segments are rendered using terrain model and geospatial database. b) Query image is segmented via semantic segmentation method. c) Semantic segments from query image are aligned with synthetic semantic segments and camera orientation is recovered.

We train the query semantic segmentation on a synthetically rendered dataset and show that synthetic data is needed to achieve reasonable accuracies when used for orientation estimation in mountainous environment. To enable matching of several semantic segment classes and an edge map with the rendered panorama, we propose a novel *confidence fusion* method which fuses individual hypotheses together to achieve better accuracy.

To estimate the camera orientation, we calculate a matching confidence by computing the cross-correlation between the query and panorama images. However, a single segment class is typically insufficient to constrain the correct rotation, as the semantic segment areas often appear similar across multiple rotations. The mutual spatial relationships between different segment classes are crucial for disambiguating the correct rotation. While a single segment class cannot resolve the angle, the combination of two segment classes produces a single distinct maximum, corresponding to the desired rotation. Assuming that the segments are accurately detected in the query image and all relevant segments are present in the rendered panorama, the correct rotation can be identified by the highest confidence across *all* fused classes. This confidence is computed by taking the product of the confidences across all classes.

Our experiments show that the proposed method outperforms state-of-the-art on publicly available test sets – GeoPose3K [Brejcha and Čadík, 2017a], Venturi Mountain dataset [Porzi et al., 2016b], and CH1 dataset [Saurer et al., 2016]. More information is available on the project website: <https://cphoto.fit.vutbr.cz/semantic-orientation/>.

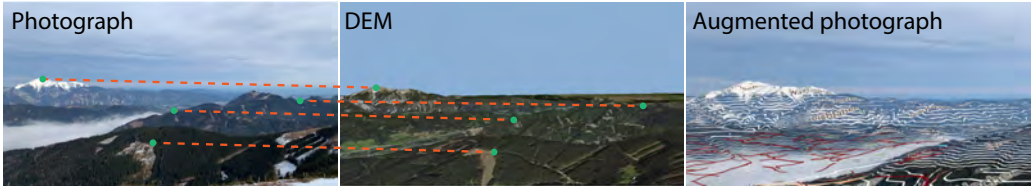


Figure 5: Our method matches a query photograph to a rendered digital elevation model (DEM). For clarity, we visualize only four matches (dashed orange). The matches produced by our system can then be used for camera pose estimation, which is a key component for augmented reality applications. In the right image (zoomed-in for clarity), we augment the view with contour lines (white), gravel roads (red), and trails (black) using the estimated camera pose.

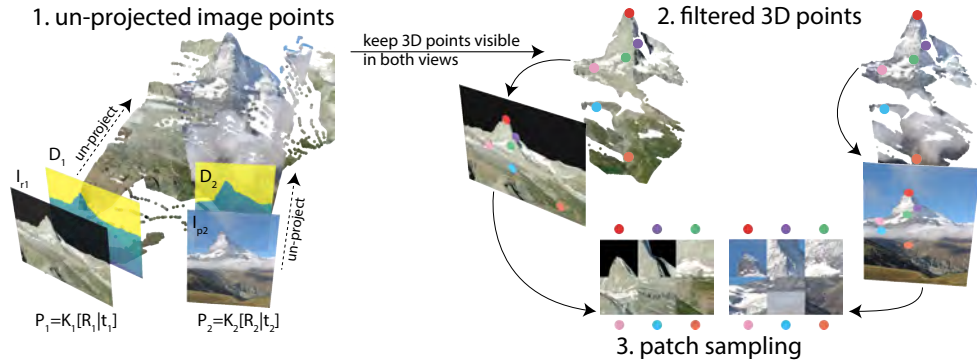


Figure 6: 1. For a pair of images (render), (photograph), 2D image points are un-projected into 3D using the rendered depth maps, and the ground truth camera poses, respectively. 2. Only points visible from both views are kept. 3. A randomly selected subset of 3D points is used to form patch centers, and corresponding patches are extracted. The patches are subsequently utilized to train the cross-domain feature descriptor.

3.2 Camera pose estimation using learned descriptors

We propose a solution for outdoor augmented reality applications by registering the user’s camera feed to large scale textured Digital Elevation Models (DEMs) [Brejcha et al., 2020]. As there is significant appearance variation between the DEM and the camera feed, we train a data driven cross-domain feature descriptor that allows us to perform efficient and accurate feature matching. Using this approach, we are able to localize photos based on long-distance cues, allowing us to display large scale augmented reality overlays such as altitude contour lines, map features (roads and trails), or 3D created content, such as educational geographic-focused features, see Fig. 5.

The key insight of our approach is that we can take advantage of a robust and readily available source of data, with near-global coverage, that is the DEM models, in order to compute camera location using reliable, 3D feature matching based methods. However, registering photographs to DEMs is challenging, as both domains are substantially different. For example, even high-quality DEMs tend to have resolution too rough to capture local high-frequency features like mountain peaks, leading to horizon mismatches. In addition, photographs have (often) unknown camera intrinsics such as focal length, exhibit seasonal and weather variations, foreground occluders like trees or people, and objects not present in the DEM itself, like buildings.



Figure 7: Camera pose estimation from line correspondences. The images are overlaid with reprojections of 3D line segments using the camera pose estimated by the our method *DLT-Combined-Lines*.

Our method works by learning a data-driven cross-domain feature embedding. We first use Structure From Motion [Snavely et al., 2008] (SfM) to reconstruct a robust 3D model from internet photographs, aligning it to a known terrain model. We then render views at similar poses as photographs, which lets us extract cross-domain patches in correspondence, see Fig. 6, which we use as supervision for training. At test time, no 3D reconstruction is needed, and features from the query image can be matched directly to renderings of the DEM. To compute full camera pose of the photograph with respect to the DEM, the *EPnP* [Lepetit et al., 2009] algorithm with RANSAC is used.

Our method is efficient enough to run on a mobile device. As a demonstration, we developed a mobile application that performs large-scale visual localization to landscape features locally on a recent iPhone, and show that our approach can be used to refine localization when embedded device sensors are inaccurate. More information is available on the project website: <https://cphoto.fit.vutbr.cz/LandscapeAR/>.

3.3 Camera pose estimation from line correspondences

We introduced novel methods for camera pose estimation based on correspondences between 3D and 2D lines, a problem commonly referred to as the Perspective- n -Line (PnL) problem. Our focus was on efficiently solving the PnL problem for large line sets using a linear formulation.

Our proposed method, *DLT-Combined-Lines* [Přibyl et al., 2017], is built on a linear formulation of the PnL problem. It combines the *DLT-Lines* method of [Hartley and Zisserman, 2004], which represents the 3D structure with 3D points, and our *DLT-Plücker-Lines* method [Přibyl et al., 2015], which represents the 3D structure using 3D lines parameterized by Plücker coordinates. This hybrid approach leverages the redundant representation of the 3D structure using both 3D points and 3D lines, reducing the minimum required line correspondences to just five.

A key element of the proposed method is the combined projection matrix recovered by the *DLT* algorithm. This matrix encapsulates multiple estimates of camera orientation and translation, enabling more accurate final camera pose estimation. The *DLT-Combined-Lines* method achieves state-of-the-art accuracy for large line sets, even under significant image noise, and performs comparably to state-of-the-art methods on real-world data (see Fig. 7). Additionally, the method retains the common advantage of LPnL approaches: it is computationally efficient and very fast.

3.4 Camera orientation estimation using cascaded attention

The advent of data-driven algorithms, particularly deep learning, has provided a new perspective on the problem of camera orientation estimation. This approach allows for the estimation of relative rotation between images, even when the images do not necessarily overlap, see Fig. 8. We proposed a novel method [Dekel et al., 2024] that directly regresses the relative rotation from input images through a deep neural network. Inspired by recent successful applications of Transformers [Vaswani et al., 2017] in computer vision tasks, we adapt Transformers for *multiple* tasks within the proposed pipeline expanding beyond previous applications of Transformers.



Figure 8: Relative camera rotation estimation from two images. The query and panoramic image is marked with green and yellow dots, respectively. The predicted rotation is represented as a footprint of the query image and it is marked by the red-dotted line. We show the results of the matching of images with large (left), small (middle) overlap and non-overlapping (right) images.

First, we apply Transformers-Decoders to improve the input image embeddings by distilling inter-image information between the images by cross-decoding, where each embedding uses the other’s embedding as a query, see Fig. 9. This better encodes images with respect to each other. Second, a Transformer-Encoder computes a stacked multihead attention to encode *cross-attention* between the latent representations of image pairs. The proposed Transformer-Encoder-based approach leverages multi-head attention’s advanced architecture to better encode interactions between activation map entries. Third, we further improve the cross-attention encoding using a cascade of two decoders and a learnt rotation query, to jointly refine the cross-attention encoding and the rotation query. The proposed scheme is a *general-purpose* attention-based architecture for estimating attributes related to two input images such as optical flow, registration, relative pose regression, etc.

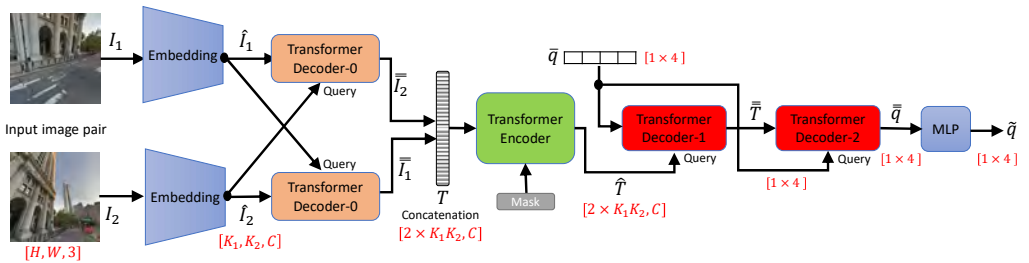


Figure 9: The proposed architecture utilizes weight-sharing Siamese CNNs to encode the input image pair $(I_1, I_2) \in \mathbb{R}^{H \times W}$ into feature maps (\hat{I}_1, \hat{I}_2) . These feature maps are then cross-decoded by the weight sharing Transformer Decoder-0 layers, cross-distilling (\hat{I}_1, \hat{I}_2) into the representations \tilde{I}_1 and \tilde{I}_2 . The concatenated refined embeddings T are input to the Transformer-Encoder alongside an attention mask M to derive the cross-attention encoding \hat{T} . \hat{T} enters a cascade of two Transformer Decoders, where the first, Transformer Decoder-1, enhances the cross-attention as \tilde{T} , guided by the learned quaternion rotation query \tilde{q} . The second, Transformer Decoder-2, encodes the rotation as \bar{q} , transformed via a multilayer perceptron (MLP) to predict the relative quaternion rotation \tilde{q} .

Interestingly, the attention maps computed by our scheme show that the Transformer-Encoder assigns high attention scores to image regions containing rotation-informative image cues, emphasizing vertical and horizontal lines. We also observe that the proposed approach can predict the rotation of non-overlapping image pairs with state-of-the-art accuracy. Our framework is end-to-end trainable and optimizes a regression loss.

4 Visual geo-localization

In this section, we present our work on visual geo-localization. To enable the evaluation of geo-localization methods and facilitate the training of deep learning models, we introduced three novel datasets [Čadík et al., 2015, Brejcha and Čadík, 2017a, Tomešek et al., 2022]. Additionally, we show our contribution to skyline (horizon line) detection [Ahmad et al., 2017, Ahmad et al., 2021], a crucial component of outdoor geo-localization techniques. Finally, we highlight our two novel approaches to image geo-localization in natural environments [Brejcha et al., 2018, Tomešek et al., 2022].

4.1 New datasets for visual localization

Until recently, the number of publicly available datasets for visual geo-localization and outdoor pose estimation was very limited. To the best of our knowledge, only two datasets existed for visual geo-localization (CH1 [Baatz et al., 2012], CH2 [Saurer et al., 2016]) and a single dataset focused on camera orientation estimation in mountainous regions [Porzi et al., 2014].

4.1.1 Dataset GeoPose3K

To address these limitations, we introduced a new dataset, *GeoPose3K*, which resolves three key issues present in existing datasets for outdoor camera pose estimation: 1) a limited number of images with ground truth positions, 2) the absence of full camera orientation data, and 3) the lack of metadata required for training and evaluating feature detectors and other applications in outdoor settings.

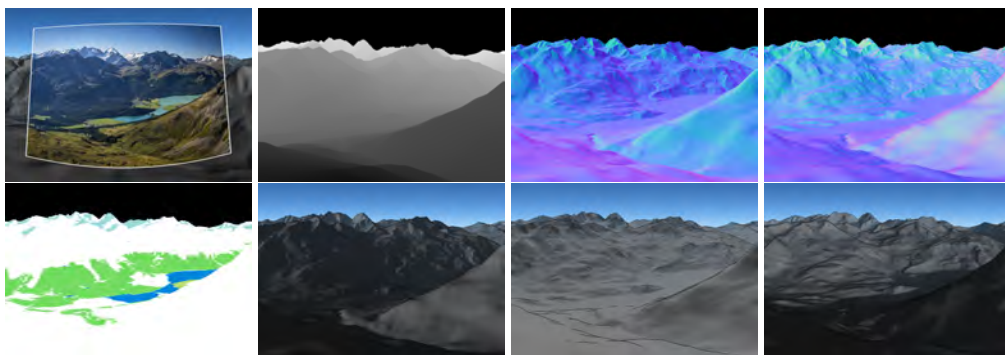


Figure 10: A sample from our *GeoPose3K* dataset: for each mountain landscape photograph, the dataset contains (in reading order) its GPS coordinate and camera orientation, distance from the camera in meters, normals w.r.t. camera, normals w.r.t. cardinal direction, semantic labels and approximate illumination during the day (here shown at 5am, 12pm and 8pm).

The *GeoPose3K* dataset (see Fig. 10) contains over three thousand photographs, primarily sourced from the photo-sharing platform Flickr.com. All images were captured in the Alps, the highest mountain range in Europe. For each image, comprehensive camera pose parameters are provided, including GPS position, field of view (FOV), and full orientation. These parameters were derived using an image-to-model matching technique and manually verified for accuracy.

To facilitate the training and development of advanced approaches for outdoor environments, we also provide a variety of synthetic data for each image, including depth maps, normal maps,

simulated illumination throughout the day, and semantic labels. An example image from our dataset, along with its corresponding synthetic data, is presented in Fig. 10.

The dataset was created using a semi-automatic method: we enhanced our existing camera orientation estimation approach [Baboud et al., 2011] by incorporating weights into the original *Alignment Metric* and training a specialized mountain silhouette detector. This improved method enabled us to develop a robust procedure for refining noisy estimations of camera position and FOV. GeoPose3K dataset is available at the project webpage <https://cphoto.fit.vutbr.cz/geoPose3K/>.

4.1.2 Dataset Alps100K

We further introduced a new dataset of 100K annotated (GPS coordinates, elevation, EXIF if available) outdoor images from mountain environments. The collection covers vast geographic area of the Alps; therefore we name it Alps100K, see Fig. 11. The images exhibit high variation in elevation as well as in landscape appearance. Furthermore, the collection spans all the seasons of the year. It contains test sets to evaluate elevation estimation performance, a large proportion of the dataset serves as a training set for the data-driven approaches.

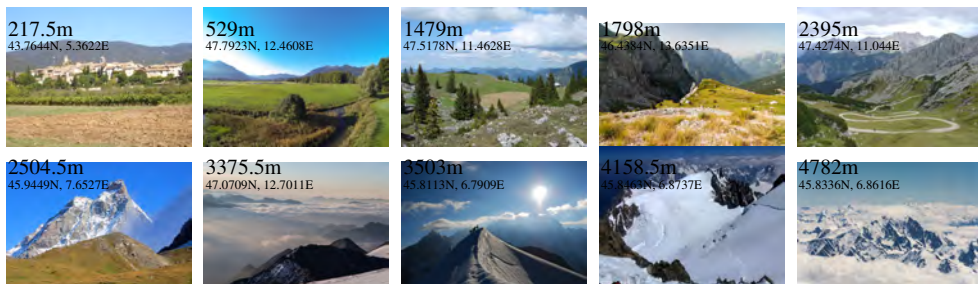


Figure 11: A sample from our dataset *Alps100K*. Each of 100K images of natural environments is accompanied with the GPS and elevation information.

Alps100K dataset acquisition: first we create a list of all hills and mountain peaks located in the seven Alpine countries (Austria, France, Germany, Italy, Liechtenstein, Slovenia, Switzerland) from the OpenStreetMap database [Curran et al., 2012]. The list of hill names is used to query the Flickr photo hosting service. In order to increase the ratio of outdoor images certain tags, such as wedding, family, indoor, still life, are excluded. Only images containing information about the camera location are kept. Out of 1.2M crawled images, about 400K are unique and inside the Alps region. To cull irrelevant (non-landscape) images a state-of-the-art scene classifier [Zhou et al., 2014] is applied. Probabilities of 205 scene categories are assigned to each image. We experimentally select 28 categories (mountain, valley, snowfield, etc.) and keep only those images whose cumulative probability in those categories exceeds 0.5. This step significantly improves the relevance of the dataset at the expense of reducing the number of images to circa 25%.

Finally the elevation of the camera is inferred from the GPS coordinates via the 3D digital elevation model¹. This model covers the Alps with 24 meter spaced samples. The collection contains 98136 outdoor images that span almost all possible elevations observed in the Alps [0, 4782m]. Geographically the dataset covers virtually all the regions of the Alps with obvious concentrations in tourist spots (e.g., around Zermatt village in Switzerland). The EXIF information is available for 41364 images, which is 42% of the Alps100K dataset. The the dataset along with elevation and GPS meta-data is available at the project webpage <https://cphoto.fit.vutbr.cz/elevation/>.

¹Available from <http://www.viewfinderpanoramas.org>

4.1.3 Dataset CrossLocate

A problem of existing datasets of photographs captured in nature is that they provide photographs usable only as queries. Moreover, the images are typically few in number (hundreds or low thousands) and sparsely distributed. To enable the development of novel visual geo-localization methods in natural environments, and to complement the existing query datasets, we created *CrossLocate* dataset, which consists of two novel databases of synthetic ground-level views – spatially non-uniform *sparse* database and *uniform* database. Each view is accompanied by detailed information about its position and orientation. The *sparse database* serves for fast and simple experiments, while the *uniform database* represents the real-world scenario of localization across a large area of hundreds of thousands of square kilometers and millions of images (see Fig. 12). Each of these databases contains three rendered image modalities – *semantic segmentations*, *silhouette maps* and (*absolute*) *depth maps*. Other modalities can be derived, such as previously used horizon lines or relative depth maps, thus enabling diverse tasks.

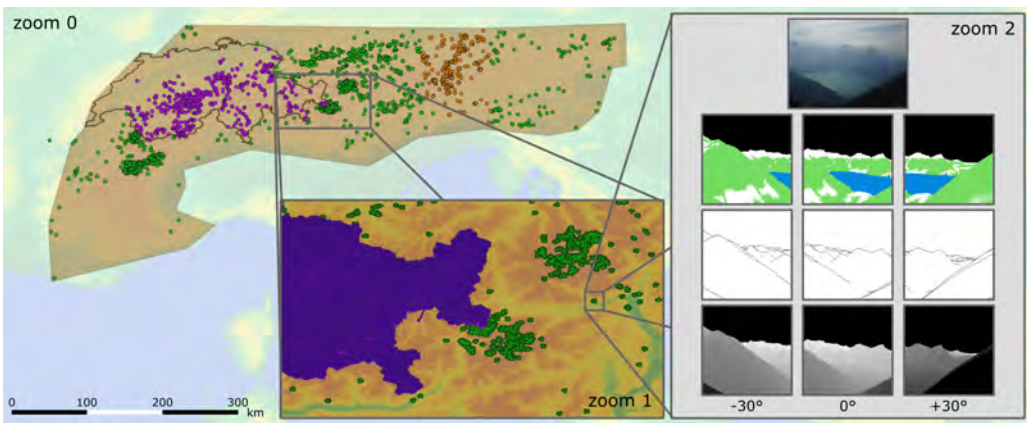


Figure 12: Our *CrossLocate* datasets at various zoom levels. *Zoom 0* shows positions of query photographs within the delimited area of the Alps. The testing area of Switzerland is marked as well. Green, orange and pink colors distinguish the proposed training, validation and testing sets, respectively. The query photographs are paired with our synthetically rendered database images. The *sparse database* contains rendered views at positions identical to the query positions. The *uniform database* has positions defined by a uniform grid with 500 m spatial sampling across the whole Alps area. *Zoom 1* provides a closer look. For simplicity of illustration, only the testing area is densely covered by the uniform grid of database positions (purple). To speed up the training process, we additionally provide the *uniform compact database* used (only) for training (dark green), where only grid positions that are within 1 kilometer from any query position are kept. *Zoom 2* offers a detailed look at a specific position, where a query photograph is complemented by rendered image modalities (semantic segmentations, silhouette maps and depth maps). There is a total of 12 views (3 shown) at each position for each modality, covering the 360° field of view.

In our terminology, a “dataset” consists of “queries” to be localized and a “database” to be searched. By combining the rendered database views with existing datasets of photographs (used as “queries” to be localized), we create a unique benchmark for visual geo-localization in natural environments, which contains correspondences between query photographs and rendered database imagery, see Fig. 12. The distinct ability to match photographs to synthetically rendered databases defines our task as “cross-modal”. Numerically, the *CrossLocate* query dataset consists of 12353

photographs and the uniform database contains over 10 million rendered images from across the Alps.

4.2 Skyline detection for visual localization

Skyline plays a pivotal role in several outdoor visual geo-localization methods [Porzi et al., 2014, Porzi et al., 2016b, Saurer et al., 2016, Baatz et al., 2012]. We present a novel mountainous skyline detection approach [Ahmad et al., 2021] where we adapt a shallow learning approach to learn a set of filters to discriminate between edges belonging to sky-mountain boundary and others coming from different regions. Unlike earlier approaches, which either rely on extraction of explicit feature descriptors and their classification, or fine-tuning general scene parsing deep networks for sky segmentation, our approach learns linear filters based on local structure analysis. At test time, for every candidate edge pixel, a single filter is chosen from the set of learned filters based on pixel’s structure tensor, and then applied to the patch around it. We then employ dynamic programming to solve the shortest path problem for the resultant multistage graph to get the sky-mountain boundary, see Fig. 13. The proposed approach is computationally faster than earlier methods while providing comparable performance and is more suitable for resource constrained platforms e.g., mobile devices, planetary rovers and UAVs.

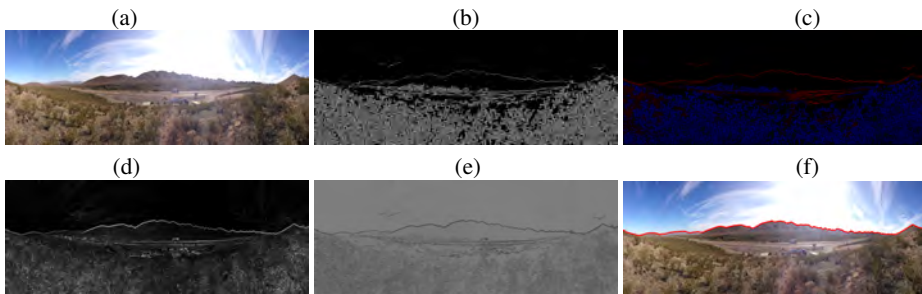


Figure 13: Our skyline detection approach: (a) input image, (b) output of the Canny edge detector, (c) predicted score for each pixel belonging to skyline using selected linear filter based on pixel’s structure tensor (brighter red and blue intensities respectively reflect more and less confidence for pixel belonging to the skyline), (d) gradient strength estimated as part of the structure tensor, (e) weighted predicted skyline score combined with gradient strength, (f) found skyline by dynamic programming overlaid on the original query image in red.

4.3 Novel geo-localization methods

We proposed two complementary approaches to image geo-localization in natural environments. The first method is designed for scenarios with sufficient coverage of community-contributed photographs [Brejcha et al., 2018]. The second proposed method relies on image retrieval from a synthetic database, allowing it to operate globally, regardless of the availability of community photographs [Tomešek et al., 2022].

4.3.1 Geo-localization using geo-registration

If the localization area is well-covered with community-contributed photographs, as is often the case for major national parks, a viable approach for geo-localization is to design a Structure-from-Motion [Snavely et al., 2008] (SfM)-based geo-localization method [Brejcha et al., 2018] as outlined below, see Fig. 14.

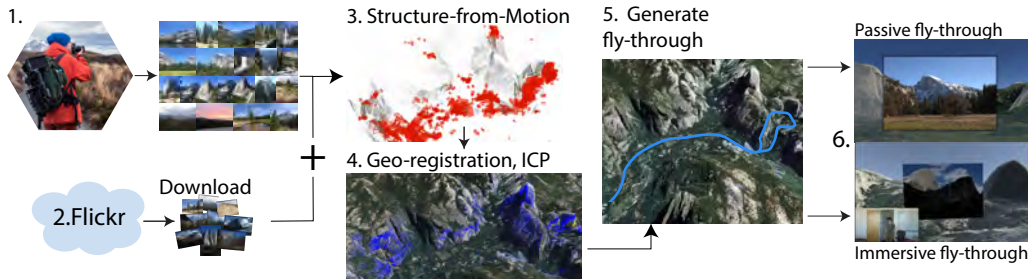


Figure 14: Visual geo-localization using structure-from-motion: 1. User takes photographs; 2. We augment the input collection with images downloaded from Flickr.com; 3. Camera positions and sparse 3D point cloud reconstruction using SfM; 4. Scene alignment with the terrain using ICP; 5. Fly-through generation from the input photographs given the geo-localization; 6. We export the fly-through to Google Earth or to our virtual reality viewer.

To enhance the coverage of the terrain and improve SfM reconstruction, we gather photographs from the specified query geo-extent by harvesting images from online repositories such as Flickr. These images are acquired using the Flickr API, with queries restricted to the specific geographic region. Furthermore, we limit the search to a defined time interval, ensuring that the downloaded images were taken during approximately the same season. This temporal constraint helps mitigate the impact of seasonal variations, thereby improving image matching and reconstruction.

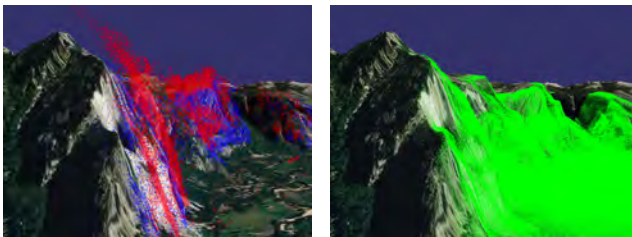


Figure 15: Alignment of input (red) point cloud with the reference (green) point cloud sampled from the terrain using Iterative Closest Points. The blue point cloud is the result.

Then, we jointly mine the photoset for both GPS metadata and visual features, which we use to obtain a rough geo-registration through a Structure from Motion (SfM) pipeline. Because of uncertainties in camera configuration, GPS location, and other parameters, there is no guarantee that the initial geo-registration actually matches the known terrain. To remedy this, we refine the initial geo-registration by minimizing the euclidean distance between the reconstructed 3D point cloud and the known DEM terrain data. We segment the point cloud into disjoint clusters so that two points in the same cluster are at most 1 km apart from each other. For each cluster, we calculate its bounding box and sample the terrain on a grid with 10 m spacing. We align the reconstructed 3D point cloud and the sampled terrain using Iterative Closest Points (ICP) [Pomerleau et al., 2013], see Fig. 15.

After registering the model, we may be left with mismatches between the photo content and the virtual terrain, most of which are due to bad information about camera configuration (e.g. focal length, exact GPS position, etc.). To correct the registration errors, we leverage our knowledge of the correspondences between 2D points observed in the photographs and the 3D points in the virtual terrain. We use these correspondences to optimize the orientation parameters using the Kabsch algorithm [Kabsch, 1976]. We project the 2D observations using camera parameters into 3D points based on the euclidean distance between camera center and the corresponding 3D point. From both sets we subtract their centroids and calculate the rotation matrix using the Kabsch algorithm.

From the geographic registration, we produce immersive presentations which are viewed either passively as a video, or interactively in virtual reality, see Fig. 14. More information is available on the project website: <https://cphoto.fit.vutbr.cz/immersive-trip-reports/>.

4.3.2 Cross modal retrieval-like geo-localization

We propose a novel image retrieval-based approach to visual geo-localization in natural environments, as illustrated in Fig. 16. Our localization method is designed as a *cross-modal image retrieval* system that operates on a synthetic database, enabling global functionality regardless of the availability of community-contributed photographs. The core idea is that, given a query image, the method identifies the most similar rendered image from the large-scale synthetic database. Because the database is geo-referenced, this retrieval enables accurate localization. Additionally, the database stores information about the camera’s yaw angle, allowing for the estimation of this angle as part of the localization process.

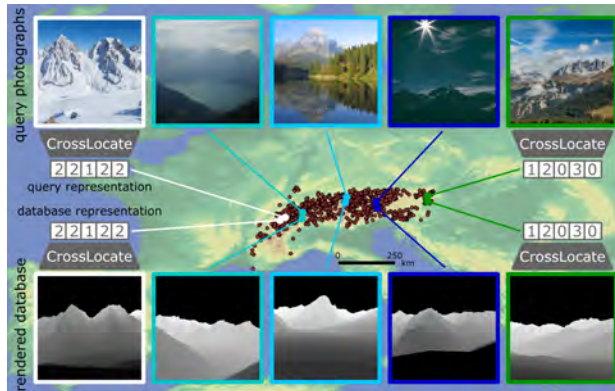


Figure 16: CrossLocate localizes ground-level photographs captured in diverse environments across the Alps. Our new databases of rendered image modalities enable the implementation of cross-modal image-retrieval (i.e., matching real photographs to rendered imagery). Out of several assessed modalities, we select depth maps and train a cross-modal global descriptor for large-scale image localization outdoors.

We aim to represent each image (place) by a single *global* descriptor. This representation is learned automatically in an end-to-end manner. Each component of our architecture is carefully selected and empirically validated. The basis for the architecture of our method are standard convolutional blocks. We use 5 convolutional blocks, with each block consisting of 2-3 convolutional layers with ReLU units, and each block is ended with max-pooling. In the last block, we do not use any pooling, and we also do not use the ReLU activation in the very last convolutional layer in order

to not restrict the resulting representation to be non-negative. Assuming a (three-channel) input image I , we obtain a 3D activation tensor $T \in R^{H \times W \times D}$ seen as $H \times W$ D -dimensional features. We apply channel-wise L2 normalization to this tensor at each of the $H \times W$ spatial positions separately.

To produce a single global descriptor as a representation for each image, we end the architecture with a global max-pooling layer, and apply another L2 normalization. The resulting descriptor has $D = 512$ dimensions. The replacement of any of these components leads to a significant drop in localization performance. We specifically stress the importance of the final max-pooling in our cross-modal task. Our single (single-branch) model is capable of extracting the deep representations for both query RGB photographs and rendered database views. When working with multiple database modalities, each modality takes one input channel.

We initialize our architecture with weights pretrained on the ImageNet dataset [Deng et al., 2009] to better cope with the low number of available images and to combat overfitting. We work with input images scaled to a unified resolution of 500×500 pixels. This resolution corresponds to the 60° field of view covered by the database views. Therefore, we scale the content of each query photograph according to its actual field of view. For example, the useful content of a query photograph with 30° field of view would be 225×225 pixels. In this way, we preserve the correct scale of the scene and enable precise localization. Taking different scales into consideration is one of the key aspects that differentiates localization from general retrieval.

We train our method using a variant of the triplet loss objective [Schroff et al., 2015], similar to [Arandjelovic et al., 2016], i.e., training is done by presenting the method with triplets of so-called query (anchor) images together with corresponding positive and negative examples. This loss function is combined with the euclidean metric, which measures the distance between extracted image representations. The goal is to learn a representation where the distances between a query and its positive example(s) are smaller than the distances between the query and its negative examples.

We further provide thorough ablation studies analyzing the localization potential of the synthetic images (modalities) rendered from a 3D terrain model, see Fig. 12. These include semantic segmentations, silhouette maps and depth maps. We reveal the depth information as the best choice for outdoor localization. Our results disclose a large gap between operating within a single image domain (e.g., photographs) and working across domains (e.g., photographs matched to rendered images), as gained knowledge is not transferable between the two. Moreover, we show that modern localization methods fail when applied to such a cross-modal task and that our method achieves significantly better results than state-of-the-art approaches. The datasets, code and trained models are available on the project website: <https://cphoto.fit.vutbr.cz/crosslocate/>.

5 Applications

In this section, we present our work on the applications of visual localization and camera orientation. Specifically, we focus on the field of automatic label placement [Bobák et al., 2019, Bobák et al., 2020, Bobák et al., 2023], which enables the automated annotation of imagery for which the camera pose has been accurately assessed. Building upon the assumption of a correctly determined camera pose, we further demonstrate how to synthesize a depth map of an image using a 3D terrain model [Čadík et al., 2018]. Finally, we present a neural long-range monocular depth estimator designed to predict absolute distances [Polasek et al., 2023]. This advancement was made possible thanks to our unique datasets [Brejcha and Čadík, 2017a, Brejcha et al., 2020], which played a crucial role in the development, training and validation of the method.

5.1 Automatic label placement

Short textual annotations (*labels*), are often used to communicate the position of features within a scene together with additional information (e.g., the name of the feature), see Fig. 17. The correct visual correspondence of the label with the annotated feature is crucial for functional and aesthetic label placement. All the labels of a single visualization form a *labeling*. In high-quality labeling, the labels should be unambiguous and well-readable, labels should not overlap with each other, one should be able to conclusively assign each annotated feature to a corresponding label and vice versa. The labeling should also be aesthetic, even though aesthetic aspects are often subjective.

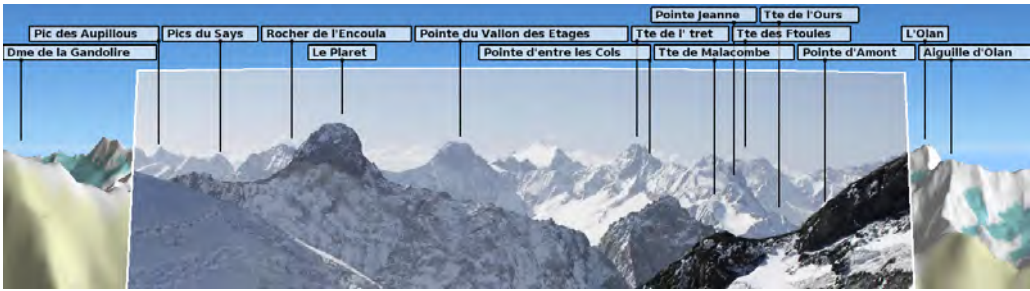


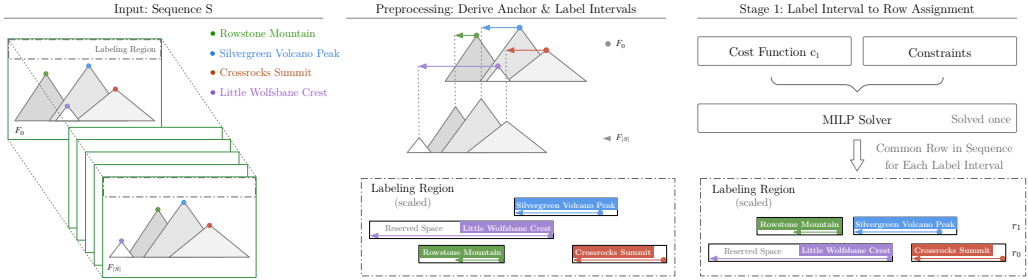
Figure 17: Automatic labeling generated by our labeling method on a photograph aligned with a geo-referenced 3D terrain model using camera pose estimation techniques [Brejcha and Čadík, 2018].

Labeling has several variants according to the type of the feature (point, area, line) and positioning of corresponding labels in the label layout. Labels can be placed within the space of visualization tightly next to the features, so-called *internal labeling*. If the feature density is too high or if the background must not be cluttered, the labels can be placed outside the visualization, so-called *external labeling*. In order to visualize the mapping of labels with corresponding features, each label is connected with its feature using a line, also known as *leader* [Bekos et al., 2006, Oeltze-Jafra and Preim, 2014].

Interactive applications (e.g., augmented reality) of labeling algorithms introduce a new aspect of temporal coherence. Applying only static algorithms on a frame-by-frame basis leads to temporally unstable behavior. In such a case, labels often jump abruptly from one position to another, breaking several assumptions of high-quality labeling.

In this work, we focus on the external labeling of dynamic scenes, where labels are placed on the top of the scene; the static case is also known as *panorama* labeling [Gemsa et al., 2014]). More specifically, we propose two temporally stable screen-space labeling methods for dynamic

OFFLINETEMPORAL Method



ONLINETEMPORAL Method

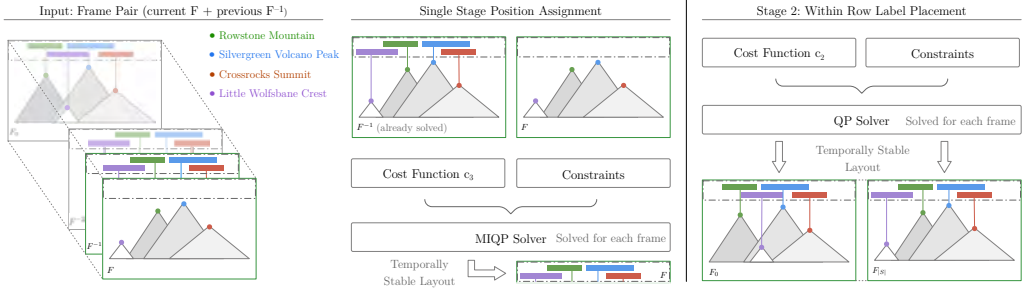


Figure 18: High-level overview of processing stages applied in the proposed labeling methods. The OFFLINETEMPORAL method consists of the *preprocessing* stage followed by the *Label Interval to Row Assignment* and *Within Row Label Placement* stage. The ONLINETEMPORAL is designed as a single-stage method.

scenes [Bobák et al., 2019, Bobák et al., 2020]. The first method called *OfflineTemporal* (see Fig. 18, top) is designated for the offline processing of the entire interaction in advance and works as follows. In the preprocessing stage, anchor intervals (representing the movement of features across frames) and label intervals (reserving space for label movement) are determined to ensure temporal coherence. Next, label intervals are assigned to discrete rows, minimizing vertical displacement and optimizing alignment with the anchors. Finally, within-row label placement is performed, where horizontal label positions are optimized while maintaining coherence and alignment with anchors. This method ensures that labels remain visually stable throughout the sequence, making it particularly suitable for pre-recorded videos such as educational content, television news infographics, or movies, where the entire sequence is available beforehand. Imagine video-footage from a drone flying through mountain terrain or a city, where one would like to label peaks or tourist attractions, respectively.

The second method, called *OnlineTemporal* (see Fig. 18, bottom), is designed for interactive applications, where frames are processed sequentially as they are generated. Unlike the multi-stage OfflineTemporal method, OnlineTemporal operates in a single stage. In this stage, the label position for each frame is calculated immediately, optimizing temporal stability by minimizing abrupt vertical and horizontal movement. Additionally, label movement between frames is restricted to a maximum of one row, ensuring smooth transitions and improving user experience during interactions. This method is particularly well-suited for real-time applications, such as games, 3D map viewers, or augmented reality, where users interact with dynamic scenes interactively. Imagine an interactive application presenting a 3D map (digital elevation model), where one would like to know nearby points of interest and could move along the scene by interacting with the camera (e.g., pan or rotate),



(a) Proposed ONLINETEMPORAL method



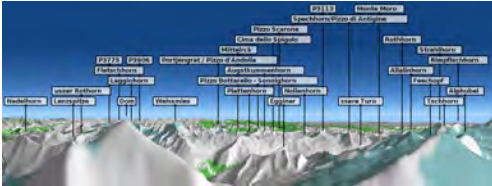
(b) Proposed ONLINETEMPORAL method with extensions



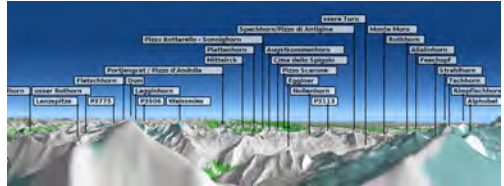
(c) Proposed OFFLINETEMPORAL method



(d) INTERVALSLOT method [Maass and Döllner, 2006]



(e) GROWINGBORDER method [Maass and Döllner, 2006]



(f) GEMSAMINROW method [Gemsa et al., 2014]

Figure 19: Example of label layouts calculated for the mountain peaks in the test sequence approximated by point anchors. Our methods (a, b, c) prioritize temporal stability and aesthetic label placement while maintaining a compact layout.

see Fig. 18.

According to the results of quantitative evaluation (see example results in Fig. 19), our labeling is more stable during the interaction than labeling produced by the current state of the art. Moreover, participants of a comprehensive user study declared that the labeling produced by the proposed methods allows them to follow moving labels significantly more accurately, and it is significantly more pleasing than with previously published methods. Furthermore, the proposed methods can be extended by the prominence of the features and easily parameterized to fit different requirements to the label layout.

5.2 Monocular depth estimation

Computational photography applications and image enhancement tasks can greatly benefit from depth information; however, estimation of outdoor depth maps remains challenging due to the vast object distances. To address this, we propose a fully automated framework for model-based generation of outdoor depth maps [Čadík et al., 2018], and a deep neural long-range monocular depth estimator [Polasek et al., 2023].

5.2.1 3D model-based synthesis of outdoor depth maps

Our first approach to depth estimation [Čadík et al., 2018] utilizes 3D terrain models and camera pose estimation techniques to generate approximate depth maps without requiring manual alignment, see Fig. 20. To overcome potential local misalignments caused by insufficient model detail or coarse registrations, we introduce a novel free-form warping method. This process begins with

the alignment of synthetic depth edges to photo edges using as-rigid-as-possible image registration, followed by further refinement of edge shapes through tight trimap-based alpha matting, see Fig. 21.

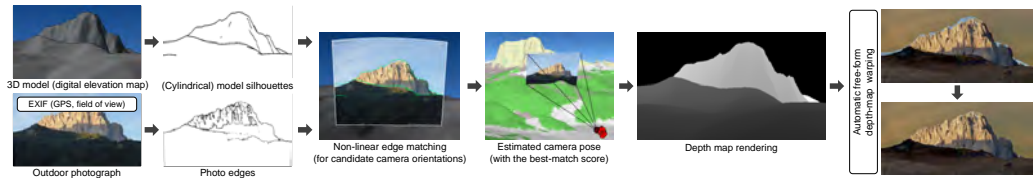


Figure 20: Overview of our fully automatic depth-map generation framework from a single landscape photograph. The camera pose to render the 3D terrain model is automatically aligned with the image. Then, the initial coarse depth map is rendered from the model using the estimated camera pose; some inaccuracies may show up due to insufficient precision of the model or due to camera alignment errors. The final depth map is refined using the free-form warping to match the local features of the input photograph.

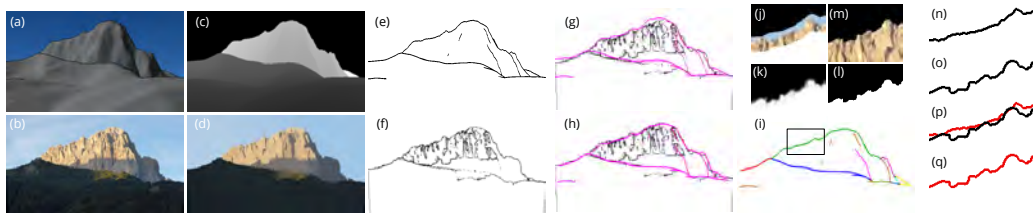


Figure 21: Aligning model depth map with the input photograph—the 3D model (a) is roughly aligned with an input photo (b), the depth map (c) is extracted using the estimated camera location and an intrinsic image [Bi et al., 2015] of the input photo (d) is computed. Depth discontinuities are extracted in the depth map (e) and edges are detected in the intrinsic image (f). Initially, model edges are misaligned with respect to photo edges (g), to reduce this misalignment as-rigid-as-possible image registration [Sýkora et al., 2009] is used (h), then edges are subdivided into individual segments (i) and tight trimap is constructed for each segment (j), alpha matte [Levin et al., 2008] is computed (k) and thresholded (l) to get the final refined shape of the photo edge (m). Given the initial model edge (n) and the refined photo edge (o), deformable image registration [Glocker et al., 2008] is used (p) to obtain the final sub-pixel accurate alignment (q).

The resulting synthetic depth maps are not only accurate but also calibrated to absolute distances. We demonstrate the effectiveness of these depth maps in several image enhancement tasks, including reblurring, depth-of-field simulation (see Fig. 22), haze removal (see Fig. 23), and guided texture synthesis, showcasing their practical value in computational photography and related fields.

5.2.2 Neural synthesis of outdoor depth maps

In our research, we also investigated neural approaches for estimating outdoor depth maps from a single image. Specifically, we developed Vision UFormer (ViUT) [Polasek et al., 2023], a deep neural network designed for long-range monocular depth estimation. ViUT consists of a Transformer [Vaswani et al., 2017, Dosovitskiy et al., 2021] encoder and a ResNet [He et al., 2016] decoder combined with the UNet [Ronneberger et al., 2015] style of skip connections, see Fig. 24. It is trained on 1M images across ten datasets in a staged regime that starts with easier-to-predict

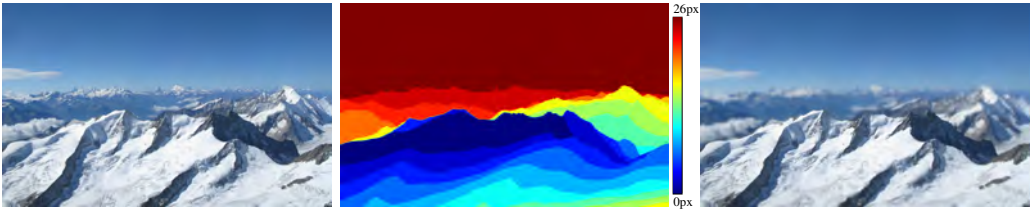


Figure 22: Transforming an outdoor photograph into a model-like look. An automatically generated synthetic depth map is used to calculate plausible blur kernel size map (middle) to simulate shallow depth-of-field (right) in landscape images (left), where such an effect cannot be achieved using standard optics for physical reasons. Virtual camera: full-frame, f-number=1.0, focal length=1200mm, focus distance=5km.

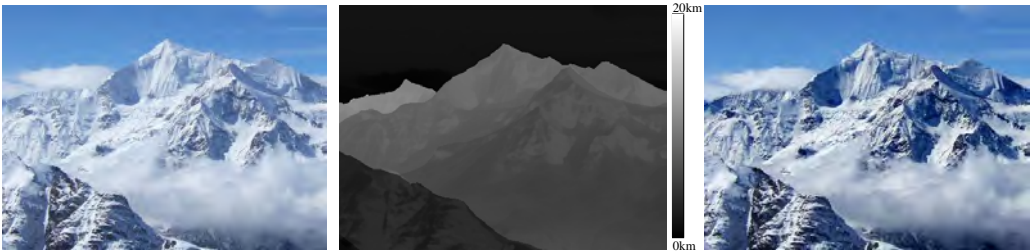


Figure 23: Example of the single-image haze removal (right) for the input photo (left) using synthetic depth map (middle). Notice the spiky peak on the horizon, which has been completely obscured by clouds in the input photo.

data such as indoor photographs and continues to more complex long-range outdoor scenes. We show that ViUT provides comparable results for normalized relative distances and short-range classical datasets such as NYUv2 [Silberman et al., 2012] and KITTI [Geiger et al., 2012]. We further show that it successfully estimates absolute long-range depth in meters. We validate ViUT on a wide variety of long-range scenes showing its high estimation capabilities with a relative improvement of up to 23%. We show its usability in image composition, range annotation, defocus, and scene reconstruction, see Figs. 25, 26.

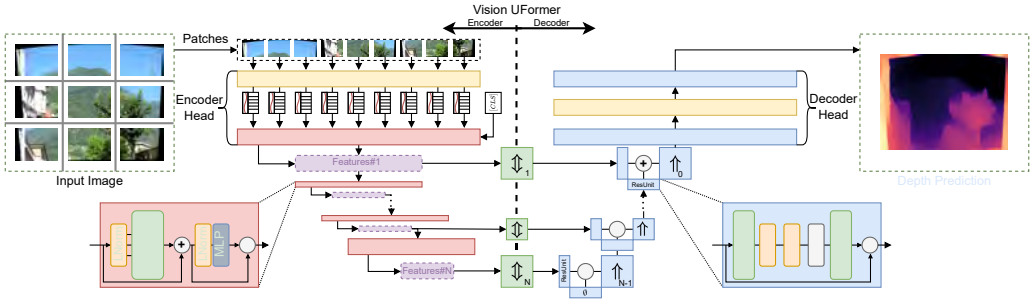


Figure 24: Vision UFormer (ViUT) overview: the model consists of a Vision Transformer [Vaswani et al., 2017, Dosovitskiy et al., 2021] encoder and ResNet [He et al., 2016] decoder in a UNet [Ronnerberger et al., 2015] configuration. The input image is split into embedded patches and passed through a sequence of multi-head self-attention layers. Multi-scale feature vectors are extracted from several tiers of the encoder and further processed into 2D feature maps. The decoder then aggregates these maps, upscaling them into the final depth prediction.



Figure 25: ViUT depth estimation: our model uses input RGB image (a) to estimate its absolute depth (b). The resulting map can then be used for further applications, such as object removal (c), 3D scene manipulation (d), or scene reconstruction (e).

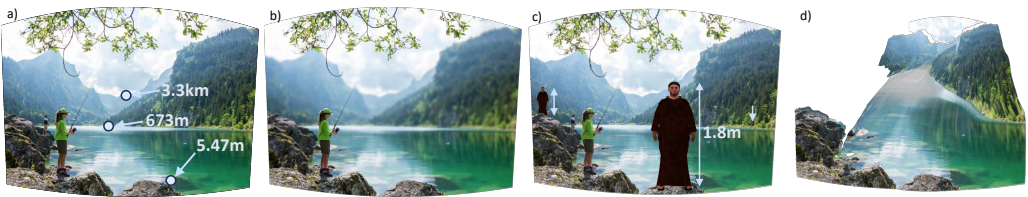


Figure 26: ViUT depth applications: we show further applications of the depths produced by our model. (a) shows direct use of depth in annotation or range-finding. We apply a depth-dependent Gaussian filter in (b) for depth of field synthesis. In (c), we show metric object insertion by placing three identical copies of a human character of approximately 1.8m in height into the scene. Thanks to the absolute depth, we observe the size of the character diminishing the further it is placed. Lastly, (d) shows a full 3D scene reconstruction with correct absolute scaling.

6 Conclusions and future work

In this thesis, we presented comprehensive research on visual geo-localization and camera pose estimation, focusing on natural environments. These settings introduce unique challenges, including sparse datasets, self-similar features, and dynamic environmental conditions. Our work addressed these challenges through innovative methods and datasets, advancing the state-of-the-art in this domain.

Key contributions include the development of novel techniques for camera orientation estimation, such as silhouette edge matching, semantic-based alignment, and learned feature descriptors. We also proposed efficient approaches for camera pose estimation using line correspondences and introduced a Transformer-based model for relative orientation estimation. These methods achieve state-of-the-art accuracy while maintaining computational efficiency.

The introduction of the GeoPose3K, Alps100K, and CrossLocate datasets represents a significant advancement in training and evaluating geo-localization methods in natural environments. These datasets provide high-quality, annotated data for a wide range of tasks, from image depth estimation to cross-modal image retrieval.

We further contributed by proposing two complementary approaches to image geo-localization in natural environments. The first method is specifically designed for scenarios with extensive coverage of community-contributed photographs, ensuring high localization accuracy in such regions. In contrast, the second approach, which relies on image retrieval from a synthetic database, offers a groundbreaking solution by enabling global operation, irrespective of the availability of community photographs.

Finally, we presented our work on the applications of visual localization and camera orientation. Our research focused on automatic label placement, depth map synthesis using a 3D terrain model, and neural long-range monocular depth estimation.

Future research directions include enhancing scalability for planet-scale localization, improving robustness under extreme environmental variations, and expanding the applicability of our methods to additional domains. The integration of advanced machine learning techniques and the use of richer datasets will further push the boundaries of visual geo-localization and camera pose estimation.

References

- [Agarwal et al., 2009] Agarwal, S., Snavely, N., Simon, I., Seitz, S. M., and Szeliski, R. (2009). Building Rome in a day. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 72–79.
- [Ahmad et al., 2017] Ahmad, T., Campr, P., Čadík, M., and Bebis, G. (2017). Comparison of semantic segmentation approaches for horizon/sky line detection. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 4436–4443.
- [Ahmad et al., 2021] Ahmad, T., Emami, E., Čadík, M., and Bebis, G. (2021). Resource efficient mountainous skyline extraction using shallow learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9.
- [Arandjelovic et al., 2016] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2016). NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307. Washington, D.C., USA: IEEE Computer Society Press.
- [Armagan et al., 2017] Armagan, A., Hirzer, M., Roth, P. M., and Lepetit, V. (2017). Learning to align semantic segmentation and 2.5D maps for geolocalization. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4590–4597. IEEE.
- [Arroyo et al., 2016] Arroyo, R., Alcantarilla, P. F., Bergasa, L. M., and Romera, E. (2016). Fusion and binarization of CNN features for robust topological localization across seasons. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4656–4663. IEEE.
- [Aydin et al., 2010] Aydin, T. O., Čadík, M., Myszkowski, K., and Seidel, H.-P. (2010). Video quality assessment for computer graphics applications. In *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia)*, pages 1–10, Seoul, Korea. ACM.
- [Aydin et al., 2010] Aydin, T. O., Čadík, M., Myszkowski, K., and Seidel, H.-P. (2010). Visually significant edges. *ACM Trans. Appl. Percept.*, 7(4):1–15.
- [Baatz et al., 2012] Baatz, G., Saurer, O., Koser, K., and Pollefeys, M. (2012). Large scale visual geolocalization of images in mountainous terrain. In *European Conference on Computer Vision (ECCV)*.
- [Baboud et al., 2011] Baboud, L., Čadík, M., Eisemann, E., and Seidel, H.-P. (2011). Automatic photo-to-terrain alignment for the annotation of mountain pictures. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '11*, pages 41–48, Washington, DC, USA. IEEE Computer Society.
- [Bae et al., 2010] Bae, S., Agarwala, A., and Durand, F. (2010). Computational rephotography. *ACM Trans. Graph.*, 29(3).
- [Bekos et al., 2006] Bekos, M. A., Kaufmann, M., Potika, K., and Symvonis, A. (2006). Multi-stack boundary labeling problems. *WSEAS Transactions on Computers*, 5(11):2602–2607.
- [Benbihi et al., 2020] Benbihi, A., Arravechia, S., Geist, M., and Pradalier, C. (2020). Image-based place recognition on bucolic environment across seasons from semantic edge description. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3032–3038.
- [Bi et al., 2015] Bi, S., Han, X., and Yu, Y. (2015). An L_1 Image Transform for Edge-Preserving Smoothing and Scene-Level Intrinsic Decomposition. *ACM Trans. Graph.*, 34(4):78.
- [Bobák et al., 2019] Bobák, P., Čmolík, L., and Čadík, M. (2019). Video sequence boundary labeling with temporal coherence. In Gavrilova, M., Chang, J., Thalmann, N. M., Hitzer, E., and Ishikawa, H., editors, *Proceedings of Computer Graphics International 2019, CGI 2019*, pages 40–52, Cham. Springer International Publishing.
- [Bobák et al., 2020] Bobák, P., Čmolík, L., and Čadík, M. (2020). Temporally stable boundary labeling for interactive and non-interactive dynamic scenes. *Computers & Graphics*, 91:265 – 278.

- [Bobák et al., 2023] Bobák, P., Čmolík, L., and Čadík, M. (2023). Reinforced labels: Multi-agent deep reinforcement learning for point-feature label placement. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–14.
- [Brachmann and Rother, 2018] Brachmann, E. and Rother, C. (2018). Learning Less is More - 6D Camera Localization via 3D Surface Regression. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4654–4662. IEEE.
- [Brahmbhatt et al., 2018] Brahmbhatt, S., Gu, J., Kim, K., Hays, J., and Kautz, J. (2018). Geometry-Aware Learning of Maps for Camera Localization. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2616–2625. IEEE.
- [Brejcha and Čadík, 2017a] Brejcha, J. and Čadík, M. (2017a). Geopose3K: Mountain landscape dataset for camera pose estimation in outdoor environments. *Image and Vision Computing*, 66:1–14.
- [Brejcha and Čadík, 2017b] Brejcha, J. and Čadík, M. (2017b). State-of-the-art in visual geo-localization. *Pattern Analysis and Applications*, 20(3):613–637.
- [Brejcha and Čadík, 2018] Brejcha, J. and Čadík, M. (2018). Camera orientation estimation in natural scenes using semantic cues. In *2018 International Conference on 3D Vision (3DV)*, pages 208–217.
- [Brejcha et al., 2018] Brejcha, J., Lukáč, M., Chen, Z., DiVerdi, S., and Čadík, M. (2018). Immersive trip reports. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, UIST '18, page 389–401, New York, NY, USA. Association for Computing Machinery.
- [Brejcha et al., 2020] Brejcha, J., Lukáč, M., Hold-Geoffroy, Y., Wang, O., and Čadík, M. (2020). LandscapeAR: Large scale outdoor augmented reality by matching photographs with terrain models using learned descriptors. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, pages 295–312, Cham. Springer International Publishing.
- [Čadík, 2007] Čadík, M. (2007). Perception motivated hybrid approach to tone mapping. In *WSCG (Full Papers)*, pages 129–136.
- [Čadík, 2008a] Čadík, M. (2008a). Perceptual evaluation of color-to-grayscale image conversions. *Comput. Graph. Forum*, 27(7):1745–1754.
- [Čadík, 2008b] Čadík, M. (2008b). *Perceptually Based Image Quality Assessment and Image Transformations*. Ph.D. thesis, Department of Computer Science and Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague.
- [Čadík, 2015] Čadík, M. (2015). *Computational Photography*. Habilitation thesis, Faculty of Information Technology, Brno University of Technology.
- [Čadík and Aydin, 2017] Čadík, M. and Aydin, T. O. (2017). Chapter 5 – HDR video metrics. In Chalmers, A., Campisi, P., Shirley, P., and Olaiwola, I. G., editors, *High Dynamic Range Video*, pages 111 – 127. Academic Press.
- [Čadík et al., 2011] Čadík, M., Aydin, T. O., Myszkowski, K., and Seidel, H.-P. (2011). On evaluation of video quality metrics: an HDR dataset for computer graphics applications. In Rogowitz, B. E. and Pappas, T. N., editors, *Human Vision and Electronic Imaging XVI*, volume 7865. SPIE.
- [Čadík et al., 2013] Čadík, M., Herzog, R., Mantiuk, R., Mantiuk, R., Myszkowski, K., and Seidel, H. (2013). Learning to predict localized distortions in rendered images. *Comput. Graph. Forum*, 32(7):401–410.
- [Čadík et al., 2012] Čadík, M., Herzog, R., Mantiuk, R., Myszkowski, K., and Seidel, H.-P. (2012). New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts. In *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia)*, volume 31, pages 1–10. ACM.
- [Čadík and Slavík, 2004a] Čadík, M. and Slavík, P. (2004a). Comparing image processing operators by means of the visible differences predictor. In *WSCG (Posters)*, pages 37–40.
- [Čadík and Slavík, 2004b] Čadík, M. and Slavík, P. (2004b). Evaluation of two principal approaches to objective image quality assessment. In *IV '04: Proceedings of the Information Visualisation, Eighth International Conference on (IV'04)*, pages 513–518, Washington, DC, USA. IEEE Computer Society.

- [Čadík and Slavík, 2005] Čadík, M. and Slavík, P. (2005). The naturalness of reproduced high dynamic range images. In *IV '05: Proceedings of the Ninth International Conference on Information Visualisation (IV'05)*, pages 920–925, Washington, DC, USA. IEEE Computer Society.
- [Čadík et al., 2003] Čadík, M., Slavík, P., and Příkryl, J. (2003). Experimental system for visualization of the light load. In *WSCG*.
- [Čadík et al., 2018] Čadík, M., Sýkora, D., and Lee, S. (2018). Automated Outdoor Depth-Map Generation and Alignment. *Elsevier Computers & Graphics*, 74:109–118.
- [Čadík et al., 2015] Čadík, M., Vašíček, J., Hradiš, M., Radenović, F., and Chum, O. (2015). Camera elevation estimation from a single mountain landscape photograph. In Xie, X., Jones, M. W., and Tam, G. K. L., editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 30.1–30.12. BMVA Press.
- [Čadík et al., 2006] Čadík, M., Wimmer, M., Neumann, L., and Artusi, A. (2006). Image attributes and quality for evaluation of tone mapping operators. In *Proceedings of the 14th Pacific Conference on Computer Graphics and Applications*, pages 35–44, Taipei, Taiwan. National Taiwan University Press.
- [Čadík et al., 2008] Čadík, M., Wimmer, M., Neumann, L., and Artusi, A. (2008). Evaluation of HDR tone mapping methods using essential perceptual attributes. *Computers & Graphics*, 32:330–349.
- [Cao et al., 2020] Cao, B., Araujo, A., and Sim, J. (2020). Unifying deep local and global features for image search. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*, pages 726–743, Cham. Springer International Publishing.
- [Curran et al., 2012] Curran, K., Crumliss, J., and Fisher, G. (2012). OpenStreetMap. *International Journal of Interactive Communication Systems and Technologies*, 2(1):69–78.
- [Dekel et al., 2024] Dekel, S., Keller, Y., and Čadík, M. (2024). Estimating extreme 3D image rotations using cascaded attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2588–2598.
- [Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, and Li Fei-Fei (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE.
- [Dosovitskiy et al., 2021] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Hounsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.
- [Fialka and Čadík, 2006] Fialka, O. and Čadík, M. (2006). FFT and convolution performance in image filtering on GPU. In *IV '06: Proceedings of the conference on Information Visualization*, pages 609–614, Washington, DC, USA. IEEE Computer Society.
- [Fišer et al., 2014] Fišer, J., Lukáč, M., Jamriška, O., Čadík, M., Gingold, Y., Asente, P., and Sýkora, D. (2014). Color Me Noisy: Example-based rendering of hand-colored animations with temporal noise control. *Computer Graphics Forum*, 33(4):1–10.
- [Ge et al., 2020] Ge, Y., Wang, H., Zhu, F., Zhao, R., and Li, H. (2020). Self-supervising fine-grained region similarities for large-scale image localization. In *Computer Vision – ECCV 2020*, volume 12349 of *Lecture Notes in Computer Science*, pages 369–386, Cham, Germany. Springer International Publishing.
- [Geiger et al., 2012] Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361.
- [Gemsa et al., 2014] Gemsa, A., Haurertand, J.-H., and Nöllenburg, M. (2014). Multi-row boundary-labeling algorithms for panorama images. *ACM TSAS*, 1(1):289–298.
- [Geppert et al., 2019] Geppert, M., Liu, P., Cui, Z., Pollefeys, M., and Sattler, T. (2019). Efficient 2d-3d matching for multi-camera visual localization. In *ICRA*, pages 5972–5978. IEEE.
- [Glocker et al., 2008] Glocker, B., Komodakis, N., Tziritas, G., Navab, N., and Paragios, N. (2008). Dense image registration through MRFs and efficient linear programming. *Medical Image Analysis*, 12(6):731–741.

- [Gronát et al., 2016] Gronát, P., Sivic, J., Obozinski, G., and Pajdla, T. (2016). Learning and Calibrating Per-Location Classifiers for Visual Place Recognition. *International Journal of Computer Vision*, 118(3):319–336.
- [Haider and Khalid, 2016] Haider, Z. and Khalid, S. (2016). Survey on effective GPS spoofing countermeasures. In *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, pages 573–577. IEEE.
- [Hartley and Zisserman, 2004] Hartley, R. I. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition.
- [Hays and Efros, 2008] Hays, J. and Efros, A. A. (2008). IM2GPS: Estimating geographic information from a single image. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. New York, NY, USA: IEEE.
- [Hays and Efros, 2015] Hays, J. and Efros, A. A. (2015). Large-scale image geolocalization. In *Multimodal Location Estimation of Videos and Images*, chapter Large-scale image geolocalization, pages 41–62. Springer International Publishing.
- [He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Computer Vision and Pattern Recognition (CVPR)*.
- [Heinly et al., 2015] Heinly, J., Sch, J. L., Dunn, E., and Frahm, J.-m. (2015). Reconstructing the World in Six Days. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [Herzog et al., 2012] Herzog, R., Čadík, M., Aydın, T. O., Kim, K. I., Myszkowski, K., and Seidel, H.-P. (2012). NoRM: no-reference image quality metric for realistic image synthesis. *Computer Graphics Forum*, 31(2):545–554.
- [Hu et al., 2018] Hu, S., Feng, M., Nguyen, R. M., and Lee, G. H. (2018). CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7258–7267. IEEE.
- [Irschara et al., 2009] Irschara, A., Zach, C., Frahm, J. M., and Bischof, H. (2009). From structure-from-motion point clouds to fast location recognition. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pages 2599–2606.
- [Kabsch, 1976] Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923.
- [Kendall et al., 2015] Kendall, A., Grimes, M., and Cipolla, R. (2015). PoseNet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pages 2938–2946. New York, NY, USA: IEEE.
- [Kopf et al., 2008] Kopf, J., Neubert, B., Chen, B., Cohen, M., Cohen-Or, D., Deussen, O., Uyttendaele, M., and Lischinski, D. (2008). Deep photo: model-based photograph enhancement and viewing. *ACM Trans. Graph.*, 27(5).
- [Lepetit et al., 2009] Lepetit, V., Moreno-Noguer, F., and Fua, P. (2009). EPnP: An accurate O(n) solution to the PnP problem. *International Journal of Computer Vision*.
- [Levin et al., 2008] Levin, A., Lischinski, D., and Weiss, Y. (2008). A closed-form solution to natural image matting. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(2):228–242.
- [Lin et al., 2013] Lin, T. Y., Belongie, S., and Hays, J. (2013). Cross-view image geolocalization. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898. IEEE.
- [Lin et al., 2015] Lin, T. Y., Cui, Y., Belongie, S., and Hays, J. (2015). Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5007–5015. IEEE.
- [Liu and Li, 2019] Liu, L. and Li, H. (2019). Lending orientation to neural networks for cross-view geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

- [Luo et al., 2011] Luo, J., Joshi, D., Yu, J., and Gallagher, A. (2011). Geotagging in multimedia and computer vision—a survey. *Multimedia Tools and Applications*, 51(1):187–211.
- [Maass and Döllner, 2006] Maass, S. and Döllner, J. (2006). Efficient view management for dynamic annotation placement in virtual landscapes. In *Smart Graphics*, pages 1–12. Springer.
- [Neumann et al., 2007] Neumann, L., Čadík, M., and Nemcsics, A. (2007). An efficient perception-based adaptive color to gray transformation. In *Proceedings of Computational Aesthetics 2007*, pages 73–80, Banff, Canada. Eurographics Association.
- [Noh et al., 2017] Noh, H., Araujo, A., Sim, J., Weyand, T., and Han, B. (2017). Large-Scale Image Retrieval with Attentive Deep Local Features. In *Proceedings of the 2017 IEEE International Conference on Computer Vision*, pages 3476–3485. IEEE.
- [Oeltze-Jafra and Preim, 2014] Oeltze-Jafra, S. and Preim, B. (2014). Survey of labeling techniques in medical visualizations. In *Proc. of VCBM '14*, pages 199–208. Eurographics Association.
- [Pajak et al., 2010a] Pajak, D., Čadík, M., Aydın, T. O., Myszkowski, K., and Seidel, H.-P. (2010a). Visual maladaptation in contrast domain. In Rogowitz, B. E. and Pappas, T. N., editors, *Human Vision and Electronic Imaging XV*, volume 7527, page 752710. SPIE.
- [Pajak et al., 2010b] Pajak, D., Čadík, M., Aydın, T. O., Okabe, M., Myszkowski, K., and Seidel, H.-P. (2010b). Contrast prescription for multiscale image editing. *The Visual Computer*, 26(6):739–748.
- [Polasek and Čadík, 2023] Polasek, T. and Čadík, M. (2023). Predicting photovoltaic power production using high-uncertainty weather forecasts. *Applied Energy*, 339:120989.
- [Polasek et al., 2023] Polasek, T., Čadík, M., Keller, Y., and Benes, B. (2023). Vision UFormer: Long-range monocular absolute depth estimation. *Computers & Graphics*, 111:180–189.
- [Polasek et al., 2021] Polasek, T., Hrusa, D., Benes, B., and Čadík, M. (2021). ICTree: Automatic perceptual metrics for tree models. *ACM Trans. Graph.*, 40(6).
- [Pomerleau et al., 2013] Pomerleau, F., Colas, F., Siegwart, R., and Magnenat, S. (2013). Comparing ICP Variants on Real-World Data Sets. *Autonomous Robots*, 34(3):133–148.
- [Porzi et al., 2016a] Porzi, L., Bulò, S. R., Lanz, O., Valigi, P., and Ricci, E. (2016a). An automatic image-to-DEM alignment approach for annotating mountains pictures on a smartphone. *Machine Vision and Applications*, 28(1-2/2017):101–115.
- [Porzi et al., 2016b] Porzi, L., Bulò, S. R., and Ricci, E. (2016b). A deeply-supervised deconvolutional network for horizon line detection. In *ACM Conference on Multimedial*.
- [Porzi et al., 2014] Porzi, L., Bulò, S. R., Valigi, P., Lanz, O., and Ricci, E. (2014). Learning contours for automatic annotations of mountains pictures on a smartphone. In *ACM/IEEE International Conference on Distributed Smart Cameras*.
- [Příbyl et al., 2016] Příbyl, B., Chalmers, A., Zemčík, P., Hooberman, L., and Čadík, M. (2016). Evaluation of feature point detection in high dynamic range imagery. *Journal of Visual Communication and Image Representation*, 2016(1):1–20.
- [Příbyl et al., 2015] Příbyl, B., Zemčík, P., and Čadík, M. (2015). Camera pose estimation from lines using Plücker coordinates. In Xie, X., Jones, M. W., and Tam, G. K. L., editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 45.1–45.12. BMVA Press.
- [Příbyl et al., 2017] Příbyl, B., Zemčík, P., and Čadík, M. (2017). Absolute pose estimation from line correspondences using direct linear transformation. *Computer Vision and Image Understanding*, 161:130–144.
- [Radenović et al., 2016] Radenović, F., Toliás, G., and Chum, O. (2016). CNN image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I*, pages 3–20.
- [Rajasekaran et al., 2022] Rajasekaran, S. D., Kang, H., Čadík, M., Galin, E., Guérin, E., Peytavie, A., Slavík, P., and Benes, B. (2022). PTRM: Perceived terrain realism metric. *ACM Trans. Appl. Percept.*, 19(2).

- [Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.
- [Saurer et al., 2016] Saurer, O., Baatz, G., Köser, K., Ladický, L., and Pollefeys, M. (2016). Image based geo-localization in the alps. *International Journal of Computer Vision (IJCV)*, 116(3):213–225.
- [Schroff et al., 2015] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.
- [Shi et al., 2019] Shi, Y., Liu, L., Yu, X., and Li, H. (2019). Spatial-aware feature aggregation for image based cross-view geo-localization. In *Advances in Neural Information Processing Systems 32*, pages 10090–10100. Curran Associates, Inc.
- [Silberman et al., 2012] Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from rgb-d images. In *European Conference on Computer Vision (ECCV)*, pages 746–760, Berlin, Heidelberg. Springer Berlin Heidelberg.
- [Snavely et al., 2008] Snavely, N., Seitz, S. M., and Szeliski, R. (2008). Modeling the world from Internet photo collections. *International Journal of Computer Vision*, 80(2):189–210.
- [Sýkora et al., 2011] Sýkora, D., Ben-Chen, M., Čadík, M., Whited, B., and Simmons, M. (2011). TexToons: Practical texture mapping for hand-drawn cartoon animations. In *Proceedings of International Symposium on Non-photorealistic Animation and Rendering*, pages 75–83.
- [Sýkora et al., 2009] Sýkora, D., Dingliana, J., and Collins, S. (2009). As-rigid-as-possible image registration for hand-drawn cartoon animations. In *Proc. Int. Symp. Non-photorealistic Animation and Rendering*, pages 25–33.
- [Sýkora et al., 2014] Sýkora, D., Kavan, L., Čadík, M., Jamriška, O., Jacobson, A., Whited, B., Simmons, M., and Sorkine-Hornung, O. (2014). Ink-and-Ray: Bas-relief meshes for adding global illumination effects to hand-drawn characters. *ACM Transaction on Graphics*, 33(2):16.
- [Tolias et al., 2013] Tolias, G., Avrithis, Y., and Jégou, H. (2013). To aggregate or not to aggregate: Selective match kernels for image search. In *2013 IEEE International Conference on Computer Vision*.
- [Tolias et al., 2020] Tolias, G., Jeníček, T., and Chum, O. (2020). Learning and aggregating deep local descriptors for instance-level recognition. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*.
- [Tomešek et al., 2022] Tomešek, J., Čadík, M., and Břejcha, J. (2022). CrossLocate: Cross-modal large-scale visual geo-localization in natural environments using rendered modalities. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2193–2202.
- [Torii et al., 2015] Torii, A., Sivic, J., Okutomi, M., and Pajdla, T. (2015). Visual Place Recognition with Repetitive Structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2346–2359.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Adv. in Neural Inf. Proc. Systems (NIPS)*, volume 30. Curran Associates, Inc.
- [Vo et al., 2017] Vo, N., Jacobs, N., and Hays, J. (2017). Revisiting IM2GPS in the Deep Learning Era. In *Proceedings of the 2017 IEEE International Conference on Computer Vision*, pages 2640–2649. IEEE.
- [Vo and Hays, 2016] Vo, N. N. and Hays, J. (2016). Localizing and orienting street views using overhead imagery. In *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I*. Springer International Publishing.
- [Wang et al., 2018] Wang, P., Yang, R., Cao, B., Xu, W., and Lin, Y. (2018). DeLS-3D: Deep Localization and Segmentation with a 3D Semantic Map. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5860–5869. IEEE.

- [Weyand et al., 2016] Weyand, T., Kostrikov, I., and Philbin, J. (2016). Planet - photo geolocation with convolutional neural networks. In *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 37–55. Cham: Springer International Publishing.
- [Zeisl et al., 2015] Zeisl, B., Sattler, T., and Pollefeys, M. (2015). Camera Pose Voting for Large-Scale Image-Based Localization. In *2015 IEEE International Conference on Computer Vision*, pages 2704–2712.
- [Zhou et al., 2014] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning Deep Features for Scene Recognition using Places Database. *NIPS*.

Abstract

This thesis presents my work, along with the contributions of my research group CPhoto@FIT, on the challenging task of visual geo-localization and camera pose estimation in natural environments. The primary objective is to determine the geographical position and orientation of a camera using visual information from captured images. Natural environments, characterized by irregular landscapes, self-similar features, and sparse datasets, pose unique challenges compared to more structured settings, such as urban areas.

We developed novel methods for camera orientation and pose estimation, utilizing techniques such as silhouette edge matching, semantic cues, line correspondences, and cascaded attention mechanisms. To support training and evaluation in outdoor localization tasks, we introduced three datasets—GeoPose3K, Alps100K, and CrossLocate. These datasets provide a diverse range of visual and synthetic modalities, including semantic segmentation, depth maps, and silhouette maps. Additionally, we contributed to skyline (horizon line) detection, a critical component of outdoor geo-localization techniques.

We further contributed by proposing two complementary approaches to image geo-localization in natural environments. The first method is specifically designed for scenarios with extensive coverage of community-contributed photographs, ensuring high localization accuracy in such regions. In contrast, the second approach, which relies on image retrieval from a synthetic database, offers a groundbreaking solution by enabling global operation, irrespective of the availability of community photographs.

Finally, we explored applications of visual geo-localization in areas such as augmented reality, automatic label placement, and depth map estimation. The methods presented in this work achieve state-of-the-art performance across various benchmarks and real-world scenarios.

Abstrakt

Tento dokument představuje moji práci a práci mé výzkumné skupiny CPhoto@FIT na téma vizuální geolokalizace a odhadu orientace kamery v přírodním prostředí. Hlavním cílem je určit geografickou polohu a orientaci kamery na základě vizuálních informací z daných snímků. Přírodní prostředí se vyznačuje nepravidelnou krajinou, soběpodobnými útvary a řídkými datovými sadami, což představuje náročné výzvy ve srovnání se strukturovanými oblastmi, jakými jsou např. městské aglomerace.

Spolu s mými kolegy jsme vyvinuli nové metody pro odhad orientace a pózy kamery, které jsou založeny na siluetních hranách, sémantické segmentaci, korespondenci čar a kaskádových mechanismech pozornosti. Pro podporu trénování a hodnocení metod vizuální lokalizace v přírodních prostředích jsme vytvořili tři datové sady – GeoPose3K, Alps100K a CrossLocate. Tyto datové sady obsahují širokou škálu vizuálních a syntetických modalit, včetně sémantické segmentace, hloubkových map a map siluet. Kromě toho jsme přispěli k detekci horizontu, která je klíčovou součástí technik vizuální geolokalizace v přírodních prostředích.

Důležitým výsledkem naší práce je návrh dvou komplementárních přístupů k vizuální geolokalizaci v přírodních prostředích. První metoda je navržena pro scénáře s dostatečným pokrytím komunitními fotografiemi, což zajišťuje vysokou přesnost lokalizace v těchto oblastech. Naproti tomu druhý přístup, který je založen na vyhledávání snímků v syntetické databázi, přináší průlomové řešení umožňující globální lokalizaci bez ohledu na dostupnost reálných fotografií.

Věnovali jsme se také aplikacím vizuální geolokalizace v oblastech, jako je rozšířená realita, automatické umišťování popisek a odhad hloubkových map. Metody prezentované v této práci dosahují kvalitních výsledků napříč různými benchmarky i reálnými scénáři, čímž poskytují cenné nástroje a poznatky, které posouvají oblast vizuální geolokalizace kupředu.